Inference and Prediction for High Dimensional Data via Penalized Regression and Kernel Machine Methods

Dissertation Advisor: Professor Tianxi Cai Jessica Nicole Minnier

Inference and Prediction for High Dimensional Data via Penalized Regression and Kernel Machine Methods

Abstract

Analysis of high dimensional data often seeks to identify a subset of important features and assess their effects on the outcome. Furthermore, the ultimate goal is often to build a prediction model with these features that accurately assesses risk for future subjects. Such statistical challenges arise in the study of genetic associations with health outcomes. However, accurate inference and prediction with genetic information remains challenging, in part due to the complexity in the genetic architecture of human health and disease.

A valuable approach for improving prediction models with a large number of potential predictors is to build a parsimonious model that includes only important variables. Regularized regression methods are useful, though often pose challenges for inference due to nonstandard limiting distributions or finite sample distributions that are difficult to approximate. In Chapter 1 we propose and theoretically justify a perturbation-resampling method to derive confidence regions and covariance estimates for marker effects estimated from regularized procedures with a general class of objective functions and concave penalties. Our methods outperform their asymptotic-based counterparts, even when effects are estimated as zero.

In Chapters 2 and 3 we focus on genetic risk prediction. The difficulty in accurate risk assessment with genetic studies can in part be attributed to several potential obstacles: sparsity in marker effects, a large number of weak signals, and non-linear effects. Single marker analyses often lack power to select informative markers and typically do not account for non-linearity. One approach to gain predictive power and efficiency is to group markers based on biological knowledge such genetic pathways or gene structure. In Chapter 2 we propose and theoretically justify a multi-stage method for risk assessment that imposes a naive bayes kernel machine (KM) model to estimate gene-set specific risk models, and then aggregates information across all gene-sets by adaptively estimating gene-set weights via a regularization procedure. In Chapter 3 we extend these methods to meta-analyses by introducing sampling-based weights in the KM model. This permits building risk prediction models with multiple studies that have heterogeneous sampling schemes.

Contents

Title page i

Abstract iii

Table of Contents v

Acknowledgments viii

1 A perturbation method for inference on regularized regression estimates 1

1.1 Introduction 2

1.2 Resampling procedures 5

1.2.1 Regularity Conditions 5

1.3 Simulation studies 11

1.4 Example: HIV drug resistance 15

1.5 Discussion 20

1.6 Acknowledgement 23

1.7 Appendix A: Proofs 24

1.7.1 Justification for the resampling method 24

1.7.2 Choice of thresholding values $\widehat{p}_{high}$ and $\widehat{p}_{low}$ for confidence regions 28

1.7.3 Justification of highest density region and bias estimate 29

1.8 Appendix B: Model selection 30

1.8.1 Selection of λ with Bayes Information Criterion 30

2 Risk classification with an adaptive naive nayes kernel machine model 32

2.1 Introduction 33

2.2 Naive-bayes kernel machine (NBKM) model 36

2.3 Model estimation under the NBKM model 38

2.3.1 Kernel PCA estimation for modeling the mth gene-Set 38

2.3.2 Combining multiple gene-sets for risk prediction 41

2.3.3 Improved estimation of $\boldsymbol{\gamma}$ via cross-validation 43

2.4 Numerical analyses 44

2.4.1 Type I diabetes GWAS dataset 44

2.4.2 Simulation studies 46

2.4.3 Data example 51

2.5 Discussion 53

2.6 Appendix A: Proofs 55

2.6.1 Convergence of the kernel PCA estimator of $\widehat{h}^{(m)}(\mathbf{z}^{(m)})$ 55

2.6.2 Convergence of eigenvectors within sample space 56

2.6.3 Convergence of eigenvectors, Nyström projection 58

2.6.4 Convergence of $\widehat{h}(\mathbf{z})$ 59

2.6.5 The oracle of gene-set weights 61

2.7 Appendix B. Simulation details 63

3 Genetic risk classification via kernel machine methods for meta-analyses with heterogenous sampling schemes 64

3.1 Introduction 65

3.2 Kernel Machine Methods 68
3.2.1 Logistic Kernel Machine Regression 68
3.2.2 Adaptive Naive-Bayes Kernel Machine (ANBKM) Classification Model 70
3.3 Weighted Estimation of Kernel Eigenfunctions 71
3.4 Meta-Analysis Model 73
3.4.1 Notation 73
3.4.2 Combining Estimates of Eigenfunctions Across Studies 74
3.5 Final Prediction Model 76
3.5.1 Model and Algorithm 76
3.5.2 Model Validation 77
3.6 Case Study: Rheumatoid Arthritis Risk Classification with Multiple Case-Control GWAS 78
3.6.1 Data 78
3.6.2 Approaches and Results 80
3.7 Simulation Studies 83
3.8 Discussion 84

Acknowledgments

It is due to a convergence of luck and the help and support of many others that I am able to submit this thesis and write these acknowledgments. Firstly, I cannot give enough thanks to my exemplary advisor, Tianxi Cai. I am thankful for her extraordinary guidance and unsurpassed patience and generosity. I am thankful that she is unfailingly committed to my success—always looking out for my best interests, challenging and motivating me, and giving me confidence. Not least, I thank her for the dinner parties, and by that I mean her friendship.

I would like to thank my committee members Victor DeGruttola and Xihong Lin, for their constructive comments in the past few years, for inspiring me to try to work as hard as they do, to strive for a fraction of success they've exhibited in their careers, and for helping me to see the context and meaning of our statistical work. I see how their work impacts public health directly, and it keeps me motivated and gives me perspective.

Thank you to my collaborators for guidance in topics as wide ranging as the most difficult theory to the most difficult data analysis. The opportunity to work with so many different researchers in so many different departments has completely shaped my experience as a student and my future career path. Without our collaborations our work has little purpose.

Many thanks are due to the generous and patient staff members who have helped me with all things I would have been helpless without. They also shape the welcoming environment of our department and support us all throughout our years there.

Special mentions go to Duque the cat for always being there for me, and for the kind people at Area Four in Cambridge who let me linger for hours on end as I worked to complete this thesis.

Lastly, I would not be here without the support and love of my closest friends, parents, aunts, and other family, and without the influence of my best professors. I know I could write many more dissertations on what I've learned from them.

A perturbation method for inference on regularized regression estimates

Jessica Minnier[1], Lu Tian[2], and Tianxi Cai[1]

[1]Department of Biostatistics
Harvard School of Public Health

[2]Department of Health Research & Policy
Stanford University School of Medicine

1.1 Introduction

Accurate prediction of disease outcomes is fundamental for successful disease prevention and treatment selection. Recent advancement in biological and genomic research has led to the discovery of a vast number of new markers that can potentially be used to develop molecular disease prevention and intervention strategies. For example, gene expression analyses have identified molecular subtypes that are associated with differential prognosis and response to treatment for breast cancer patients (Perou et al., 2000; Dent et al., 2007). For non-small cell lung cancer patients, a composite score consisting of several biological markers including cyclin E and Ki-67 was shown to be highly predictive of patient survival (Dosaka-Akita et al., 2001). However, construction of accurate prediction models with a panel of markers is a difficult task in general. For example, statistical models for calculating individual cancer risk have been developed for a few types of cancer in the past two decades (Gail et al., 1989; Thompson et al., 2006; Cassidy et al., 2008; Freedman et al., 2009). However, much refinement is needed even for the best of these models due to their limited discriminatory accuracy (Spiegelman et al., 1994; Gail and Costantino, 2001).

The increasing availability of new potential markers, while holding great promise for better prediction of disease outcomes, imposes challenges to model development due to the high dimensionality in the feature space and the relatively small sample size. To improve prediction with a large number of promising genomic or biological markers, an important step is to build a parsimonious model that only includes important markers. Such a model could reduce the cost associated with unnecessary marker measurements and improve the prediction precision for future patients. For such purposes, various regularization procedures such as the LASSO (Tibshi-

rani, 1996; Knight and Fu, 2000), the SCAD (Fan and Li, 2001, 2002, 2004; Zhang et al., 2006), the adaptive LASSO (ALASSO; Zou, 2006; Wang and Leng, 2007), the Elastic Net (Zou and Hastie, 2005; Zou and Zhang, 2009), and one-step local linear approximation (LLA; Zou and Li, 2008) have been developed in recent years. These procedures simultaneously identify non-informative variables and produce coefficient estimates for the selected variables to induce a model for prediction.

These regularization procedures, while effective for variable selection and stable estimation, yield estimators whose distributions are difficult to approximate. LASSO type estimators have a non-standard limiting distribution that depends on which components of the coefficient vector are zero. Since the LASSO type estimator is not consistent in variable selection, the limiting distribution cannot be estimated directly. Furthermore, standard bootstrap methods fail when the true coefficient vector is sparse (Knight and Fu, 2000). Recently, Chatterjee and Lahiri (2010) proposed a truncated LASSO estimator whose distribution can be approximated using a residual bootstrap procedure. To overcome the difficulties in LASSO estimators, other regularized procedures such as the SCAD and ALASSO have been proposed. These estimators possess asymptotic *oracle* properties including perfect variable selection and super efficiency. However, our simulation results suggest that in finite samples, such oracle properties are far from being true and inference procedures based on asymptotic properties such as those given in Zou (2006) perform poorly especially when the signal to noise ratio (SNR) is high and the between covariate correlations are not low. Recently, Pötscher and Schneider (2009, 2010) developed theory on the coverage probabilities of the confidence intervals for ALASSO type estimators under the orthogonal design. It was shown that estimating the distribution function of the ALASSO estimator is not feasible when the true parameter is of similar magnitude

to $n^{-\frac{1}{2}}$, where n is the sample size. It is thus generally difficult to develop well performed confidence regions (CRs) and hypothesis testing procedures based on these regularized estimators. Such difficulties limit applicability to clinical studies where confidence in statistical evidence is crucial for clinical decision making.

In this paper, we propose resampling methods to derive CR and testing procedures for marker effects estimated from regularized procedures such as the ALASSO and one-step SCAD estimator when the true parameter is fixed. Our preliminary studies suggest that CRs constructed from such resampling procedures perform much better than their asymptotic based counterparts. When the fitted model is merely a *working model*, many frequently used estimation procedures may fail to produce stable parameter estimates. Procedures that can provide stable parameter estimates and valid interval estimates under a possibly misspecified working model are highly valuable when building a prediction model with high dimensional data. Our proposed procedures remain valid even if the fitted model fails to hold, provided that the employed objective function satisfies mild regularity conditions. The rest of the paper is organized as follows. In Section 2, we introduce the proposed perturbation resampling procedures and describe various methods for constructing confidence regions. In Section 3, we demonstrate the validity of the proposed procedures in finite samples via simulation studies. In Section 4, we illustrate our proposed procedure with an HIV drug resistance study where the goal is to predict phenotypic drug resistance levels using genotypic viral mutations.

1.2 Resampling procedures

Suppose that $\mathbf{y} = (y_1, \ldots y_n)^\intercal$ is the $n \times 1$ vector of response variables and $\mathbf{x}_j = (x_{1i}, \ldots, x_{pi})^\intercal, i = 1 \ldots n$, are the predictors. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\intercal$ be the $n \times p$ matrix of these covariates. Assume that the effect of $\mathbf{x}$ on y is determined via an objective function $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \ell(y, \alpha + \boldsymbol{\beta}^\intercal \mathbf{x})$, where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^\intercal)^\intercal$, α is an unknown location parameter, $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of covariate effects, and $\mathcal{D} = (y, \mathbf{x}^\intercal)^\intercal$. To assess the association between $\mathbf{x}$ and y, let $\widetilde{\mathcal{L}}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_i)$ be the objective function used to fit a regression model and $\widetilde{\boldsymbol{\theta}} = (\widetilde{\alpha}, \widetilde{\boldsymbol{\beta}}^\intercal)^\intercal = \text{argmin}_{\boldsymbol{\theta}} \widetilde{\mathcal{L}}(\boldsymbol{\theta})$. To obtain a regularized estimator for $\boldsymbol{\theta}_0$, we minimize the regularized objective function

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \widetilde{\mathcal{L}}(\boldsymbol{\theta}) + \sum_{j=1}^{p} p'_{\lambda_n j}(|\widetilde{\beta}_j|)|\beta_j| \tag{1.1}$$

where $p'_{\lambda_n j}(|\widetilde{\beta}_j|)$ is the derivative of a penalty $p_{\lambda_n j}(|\beta_j|)$ evaluated at the initial estimate of β_{0j}. We consider the cases where $p_{\lambda_n j}(|\beta_j|)$ is the concave SCAD penalty or the L_q penalty for $0 < q < 1$, and utilize a one-step estimator of these penalties with the local linear approximation (LLA) method proposed by Zou and Li (2008). Additionally, we consider the ALASSO penalty of Zou (2006) that arises when $p'_{\lambda_n j}(|\widetilde{\beta}_j|) = n^{-\frac{1}{2}} \lambda_n |\widetilde{\beta}_j|^{-1}$.

1.2.1 Regularity Conditions

To ensure the asymptotic oracle properties of the regularized estimators and the validity of the proposed resampling procedures, we require the following set of conditions:

C1. $\mathbb{P}\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}$ has a unique minimum at $\boldsymbol{\theta}_0$ and a continuous secondary derivative

with a positive definite $\mathbb{A} = \partial^2 \mathbb{P}\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}/\partial\boldsymbol{\theta}\boldsymbol{\theta}^{\intercal}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} > 0$, where $\mathbb{P}$ is the probability measure generated by the data $\mathcal{X} = \{\mathcal{D}_i, i = 1, ..., n\}$.

C2. The class of functions indexed by $\boldsymbol{\theta}$, $\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) \mid \boldsymbol{\theta} \in \Omega\}$, is Glivenko-Cantelli (Kosorok, 2008), where $\mathcal{D} = (y, \mathbf{x}^T)^T$ and Ω is the compact parameter space containing $\boldsymbol{\theta}_0$.

C3. There exists a "qausi-derivative" function $\mathcal{U}(\boldsymbol{\theta}; \mathcal{D})$ for $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$ such that for any positive sequence $\delta_n \to 0$

(a) $\mathbb{P}\{\mathcal{U}^{\otimes 2}(\boldsymbol{\theta}_0; \mathcal{D})\} = \mathbb{B}$, a positive definite matrix.

(b) $\mathbb{P}\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) - \mathcal{L}(\boldsymbol{\theta}_0; \mathcal{D}) - \mathcal{U}(\boldsymbol{\theta}_0; \mathcal{D})(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\} = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbb{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2)$, where $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_n$.

(c) $\mathbb{P}_n\{\mathcal{L}(\boldsymbol{\theta}_1; \mathcal{D}) - \mathcal{L}(\boldsymbol{\theta}_2; \mathcal{D}) - \mathcal{U}(\boldsymbol{\theta}_2; \mathcal{D})(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\} = \frac{1}{2}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T \mathbb{A}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) + o(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + n^{-1/2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|)$, almost surely, uniformly in $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\| \leq \delta_n, \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_0\| \leq \delta_n$.

These conditions are parallel to the conditions required in Proposition A1-A3 in Jin et al. (2001). These regularity conditions hold for commonly used L_2 minimization with $\mathcal{L}(\boldsymbol{\beta}; \mathcal{D}) = (y - \boldsymbol{\beta}^{\intercal}\mathbf{x})^2$ and L_1 minimization with $\mathcal{L}(\boldsymbol{\beta}; \mathcal{D}) = |y - \boldsymbol{\beta}^{\intercal}\mathbf{x}|$. Details of the justification for these two cases can be found in Section 3 of Jin et al. (2001). These conditions also guarantee that $\widetilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$ and $n^{\frac{1}{2}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to $N(\mathbf{0}, \mathbb{A}^{-1}\mathbb{B}\mathbb{A}^{-1})$. Let $\mathcal{A} = \{j : \beta_{0j} \neq 0\}$ of size q and $\mathcal{A}^c = \{j : \beta_{0j} = 0\}$, where a_j denotes the jth component of a vector $\boldsymbol{a}$.

Following similar arguments to those given in Zou (2006), Zou and Li (2008) and the unconditional arguments given in the appendix, $\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \widehat{\mathcal{L}}(\boldsymbol{\theta})$ has 'good' properties for certain choices of λ_n, including the oracle property,

Lemma 1: (Oracle properties) Suppose that $\lambda_n \to 0$ and $\lambda_n n^{\frac{1}{2}} \to \infty$. Then the regularized estimates must satisfy the following:

1. Consistency in variable selection: $\lim_n \mathbb{P}\{I(\widehat{\mathcal{A}} = \mathcal{A}) = 1\} = 1$, where $\widehat{\mathcal{A}} = \{j : \hat{\beta}_j \neq 0\}$.
2. Asymptotic normality: $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0\mathcal{A}}) \to_d {}^*\mathrm{N}(\mathbf{0}, \mathbb{A}_{11}^{-1}\mathbb{B}_{11}\mathbb{A}_{11}^{-1})$, where $\mathbb{A}_{11}$ and $\mathbb{B}_{11}$ are the respective $q \times q$ submatrices of $\mathbb{A}$ and $\mathbb{B}$ corresponding to $\mathcal{A}$.

This lemma guarantees that the regularized estimate asymptotically chooses the correct model and has the optimal estimation rate. However, estimating the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ in finite samples remains difficult. To estimate the standard errors of the SCAD estimates $\widehat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}}\{\widetilde{\mathcal{L}}(\boldsymbol{\theta}) + \sum_{j=1}^{p} p_{\lambda_n j}(|\beta_j|)\}$ when $\widetilde{\mathcal{L}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_i)$ is smooth in $\boldsymbol{\theta}$, Fan and Li (2001) proposed a local quadratic approximation (LQA) method. This gives a sandwich estimator for the covariance matrix of the estimated nonzero parameters:

$$\widehat{\mathrm{cov}}\left(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}}\right) = \{\nabla^2\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}}) + \Sigma_\lambda(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}})\}^{-1}\widehat{\mathrm{cov}}\{\nabla\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}})\}\{\nabla^2\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}}) + \Sigma_\lambda(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}})\}^{-1} \quad (1.2)$$

where $\nabla\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}}) = \partial\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}})/\partial\boldsymbol{\theta}$, $\nabla^2\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}}) = \partial^2\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$, and $\Sigma_\lambda(\widehat{\boldsymbol{\theta}}_{\widehat{\mathcal{A}}})$ is a diagonal matrix with the (j,j)th element being $I(\widehat{\beta}_j \neq 0)p'_{\lambda_n 1}(|\widehat{\beta}_j|)/|\widehat{\beta}_j|$. The LQA approach can also be used to construct a covariance estimate for the ALASSO estimates where $p'_{\lambda_n j}(|\widetilde{\beta}_j|) = n^{-\frac{1}{2}}\lambda_n|\widetilde{\beta}_j|^{-1}$. Similar to covariance estimates in Tibshirani (1996) and Fan and Li (2001) for penalized estimates, this procedure estimates the standard errors for variables with $\hat{\beta}_j = 0$ as 0. Although this sandwich estimator has been proven to be consistent (Fan and Peng, 2004) under the linear regression model, it tends to underestimate the standard errors, and normal confidence regions (CRs) using this estimate often do not provide acceptable coverage in finite sample.

To approximate the covariance of $\widehat{\boldsymbol{\theta}}$ more accurately, we propose a perturbation method to estimate the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ for a general class of objective functions and penalties. Let $\mathcal{G} = \{G_i, i = 1, \ldots, n\}$ be a set of independent and identically distributed (*i.i.d.*) positive random variables with mean and variance equal to one. We first perturb the initial objective function and obtain

$$\widetilde{\mathcal{L}}^*(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_i) G_i, \qquad \text{and} \qquad \widetilde{\boldsymbol{\theta}}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \widetilde{\mathcal{L}}^*(\boldsymbol{\theta}). \tag{1.3}$$

Then with the same set $\mathcal{G}$, we obtain the minimizer of a stochastically perturbed version of the regularized objective function:

$$\widehat{\mathcal{L}}^*(\boldsymbol{\theta}) = \widetilde{\mathcal{L}}^*(\boldsymbol{\theta}) + \sum_{j=1}^{p} p'_{\lambda_n^* j}(|\widetilde{\beta}_j^*|)|\beta_j| \tag{1.4}$$

where λ_n^* satisfies the same order constraints as λ_n as discussed in the Lemma 1. In practice, one may select λ_n and λ_n^* based on the BIC criterion detailed in the appendix with the corresponding objective functions. In the appendix we first show that $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}} - \boldsymbol{\theta}_{0\mathcal{A}})$ converges in distribution to $N(\mathbf{0}, \mathbb{A}_{11}^{-1}\mathbb{B}_{11}\mathbb{A}_{11}^{-1})$, the same limiting distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Furthermore, $\mathbb{P}^*(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}^c} = 0) \to 1$, where $\mathbb{P}^*$ is the probability measure generated by both $\mathcal{X}$ and $\mathcal{G}$. In addition, we show that the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})$ conditional on the data can be used to approximate the unconditional distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{\mathcal{A}0})$ and that $\mathbb{P}^*(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}^c} = 0 \mid \mathcal{X}) \to 1$. In practice, these results allow us to estimate the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ by generating a large number, M, say, of random samples $\mathcal{G}$. We obtain $\widehat{\boldsymbol{\theta}}^*_m$ by minimizing the perturbed objective function for each sample $m = 1, \ldots M$, and then approximate the theoretical distribution of $\widehat{\boldsymbol{\theta}}$ by the empirical distribution $\{\widehat{\boldsymbol{\theta}}^*_m, m = 1, \ldots M\}$. Specifically, the covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by the sample covariance matrix constructed from $\{\widehat{\boldsymbol{\theta}}^*_m, m = 1, \ldots M\}$.

Estimating the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ based on the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*-\widehat{\boldsymbol{\theta}}) \mid \mathcal{X}$ leads to the construction of three possible $(1-\alpha)100\%$ confidence regions for $\boldsymbol{\theta}_0$. For the first, let $\widehat{\sigma}_j^2 = M^{-1}\sum_{m=1}^{M}(\widehat{\beta}_{mj}^* - \widehat{\beta}_j)^2$. We construct a normal CR for β_{0j}, $\mathrm{CR}_j^{*\mathrm{N}}$, centered at $\widehat{\beta}_j$ with standard deviation $\widehat{\sigma}_j^*$. Since $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*-\widehat{\boldsymbol{\theta}})|\mathcal{X}$ and $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ converge to the same normal distribution, $n\widehat{\sigma}_j^2$ consistently estimates the variance of $n^{\frac{1}{2}}(\widehat{\beta}_j-\beta_{0j})$. This method is in contrast to $\mathrm{CR}^{\mathrm{Asym}}$ obtained with standard deviations $\widehat{\sigma}_j^{\mathrm{Asym}}$ estimated with the asymptotically consistent LQA sandwich estimator in Fan and Li (2001) and Zou (2006). In contrast to setting the standard error to 0 when $\widehat{\beta}_j = 0$, we set $\mathrm{CR}_j^{*\mathrm{N}} = \{0\}$ if the proportion of $\widehat{\beta}_j^*$ being 0 is larger than a threshold $\widehat{p}_{high}$, such that $\widehat{p}_{high} \to p_{high} < 1$. This method accounts for the superefficiency due to the oracle property and results in a shorter interval with valid coverage. Secondly, we simply take the $(\alpha/2)100$th and $(1-\alpha/2)100$th quantiles of $\widehat{\beta}_j^*$ as the upper and lower bounds of $\mathrm{CR}_j^{*\mathrm{Q}}$. For the third, we estimate the density of $\widehat{\beta}_j^*$ with a kernel density estimator and choose the $(1-\alpha)100\%$ highest density region, $\mathrm{CR}_j^{*\mathrm{HDR}}$. We estimate the density of $\widehat{\beta}_j^* \mid \mathcal{X}$ as a mixed density with distribution $f_j^*(\beta) = \widehat{\mathcal{P}}_{0j} I(\beta = 0) + (1-\widehat{\mathcal{P}}_{0j}) f_j^*(\beta)$, where $\widehat{\mathcal{P}}_{0j}$ is the proportion of $\widehat{\beta}_j^*$ set to 0, and $f_j^*(\beta)$ is the unknown distribution of $\widehat{\beta}_j^*$ given that it is not set to 0. To construct a highest density confidence region that has accurate coverage of this mixed distribution, we adjust the definition of the region depending on thresholds that reflect the strength of evidence for $\beta_{0j} = 0$. Our highest density confidence region $\mathrm{CR}_j^{*\mathrm{HDR}}$ is defined as

$$\mathrm{CR}_j^{*\mathrm{HDR}} = \begin{cases} \{0\} & \text{if } \widehat{\mathcal{P}}_{0j} \geq \widehat{p}_{high} \\ \{\beta : f_j^*(\beta) \geq \widehat{c}_1\} \cup \{0\} & \text{if } \widehat{p}_{low} \leq \widehat{\mathcal{P}}_{0j} < \widehat{p}_{high} \\ \{\beta : f_j^*(\beta) \geq \widehat{c}_2\} \cup \{0\} & \text{if } \alpha \leq \widehat{\mathcal{P}}_{0j} < \max(\alpha, \widehat{p}_{low}) \\ \{\beta : f_j^*(\beta) \geq \widehat{c}_3\} & \text{if } \widehat{\mathcal{P}}_{0j} < \alpha \end{cases} \tag{1.5}$$

where $\widehat{c}_1$, $\widehat{c}_2$, and $\widehat{c}_3$ are chosen such that for $H(c) = \int I\{f_j^*(\beta) \geq c\} f_j^*(\beta) d\beta$, we have $H(\widehat{c}_1) = (1 - \alpha - \widehat{\mathcal{P}}_{0j})/(1 - \widehat{\mathcal{P}}_{0j}), H(\widehat{c}_2) = 1 - \alpha + \alpha(\widehat{\mathcal{P}}_{0j} + \widehat{p}_{low}), H(\widehat{c}_3) = 1 - \alpha$, while $\widehat{p}_{low} \to 0$ and $\widehat{p}_{high} \to p_{high} = 1 - \alpha$. When $\widehat{\mathcal{P}}_{0j}$, the proportion of $\widehat{\beta}_j^*$ set to zero, is greater than the upper thresholding value $\widehat{p}_{high}$, we have strong evidence that $\beta_{0j} = 0$ and thus take $\{0\}$ as the confidence interval. When $\widehat{\mathcal{P}}_{0j}$ is between the high and low thresholding $\widehat{p}$ values, we have moderately strong evidence that $\beta_{0j} = 0$ and thus take the mass at 0 and a $1 - \alpha - \widehat{\mathcal{P}}_{0j}$ highest density region from the $\widehat{\beta}_j^* \mid \widehat{\beta}_j^* \neq 0$ samples. The occurrence of $\alpha \leq \widehat{\mathcal{P}}_{0j} < \max(\alpha, \widehat{p}_{low})$ suggests that β_{0j} is likely to be a weak signal. For such cases, it would be difficult to make inference about β_{0j} due to shrinkage. Thus, we inflate the highest density region from the $\widehat{\beta}_j^* \mid \widehat{\beta}_j^* \neq 0$ samples. Finally, when $\widehat{\mathcal{P}}_{0j} < \alpha$, we have strong evidence that β_{0j} is nonzero and so we take the $1 - \alpha$ highest density region of the continuous empirical distribution of nonzero $\widehat{\beta}_j^*$ samples. The justification of this method and the choices of $\widehat{p}_{high}$ and $\widehat{p}_{low}$ are relegated to the appendix.

In practice, when assessing the effects of multiple features, it is often important to adjust for multiple comparisons. For interval estimation, we may construct a $(1 - \alpha)100\%$ simultaneous confidence region to cover the entire parameter vector $\boldsymbol{\theta}_0$. We may then make statements about the importance of each of the covariates in the presence of other covariates while maintaining a type I error of α. For the regularized estimator, we define the Normal simultaneous region as $\text{CR}^{*\text{Sim}} = \prod_{j \notin \widehat{\mathcal{A}}^*} \{0\} \times \prod_{j \in \widehat{\mathcal{A}}^*} (\hat{\beta}_j - \gamma_\alpha \hat{\sigma}_j^*, \hat{\beta}_j + \gamma_\alpha \hat{\sigma}_j^*)$ where $\widehat{\mathcal{A}}^* = \{j : \widehat{\mathcal{P}}_{0j} < \widehat{p}_{high}\}$ and γ_α is the $(1 - \alpha)100\%$ quantile of $\left\{ \max \left\{ |\hat{\beta}_{jm}^* - \hat{\beta}_j| / \widehat{\sigma}_j^* \right\}_{j \in \widehat{\mathcal{A}}^*} \right\}_{m=1}^{M}$. We define the $(1 - \alpha)100\%$ HDR simultaneous region as $\text{CR}^{*\text{SimHDR}} = \prod_j \text{CR}_{j,\alpha_s}^{*\text{HDR}}$ where $\text{CR}_{j,\alpha_s}^{*\text{HDR}}$ is the 1-α_s $\text{CR}_j^{*\text{HDR}}$ for $\widehat{\beta}_j$ and $\alpha_s = 2(1 - \Phi(\gamma_\alpha))$. We compare the performance of these confidence regions with numerical examples in Sections 1.3 and 1.4.

1.3 Simulation studies

To examine the validity of our procedures in finite samples, we performed simulation studies to assess the performance of the corresponding confidence regions. For each setting, we simulated 1500 data sets with n observations generated under the linear model, $y = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $x_{ij} \sim \mathcal{N}(0, 1)$, the pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ was set to $cor(\mathbf{x}_i, \mathbf{x}_j) = \rho$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and $\boldsymbol{\beta}, \rho$, and σ were varied between settings. In each setting, $\boldsymbol{\beta}$ was sparse and included medium and high signals. We obtained ALASSO estimators via LARS (Efron et al., 2004) for each simulated data set with OLS initial estimates and λ chosen by the BIC as described in the appendix, and then $M = 500$ perturbed samples using our proposed method with $\mathcal{G}$ generated from a mean 1 exponential distribution. The sample size n was set to 100, 200, 400, or 1000, while ρ was 0, 0.2, or 0.5, and σ was 1 or 2. To compute the highest density regions $\text{CR}^{*\text{HDR}}$ we utilized the `hdrcde` package in `R` with the "ndr" bandwidth estimator as presented in Scott (1992) based on Silverman's rule of thumb (Silverman, 1986). We chose $\widehat{p}_{low} = \min\{\sqrt{2/\pi}\exp(-n\lambda/(4\widehat{\sigma}^2)), 0.49\}$ and $\widehat{p}_{high} = \min\{1 - \sqrt{2/\pi}\exp(-n\lambda/\widehat{\sigma}^2), 0.95\}$ as justified in the appendix for $\text{CR}^{*\text{HDR}}$, $\text{CR}^{*\text{N}}$, $\text{CR}^{*\text{Sim}}$, and $\text{CR}^{*\text{SimHDR}}$. We substituted the σ used in the standard deviation estimate from Zou (2006) analogous to equation (1.2) with the known σ from the simulations. We present the results for simulations with $n = 100$, 200 and 400 when $\sigma = 1$ or 2 and $p = 10$ or 20. In these cases, the true $\boldsymbol{\beta}_0$ contains two large effects of $\beta_{0j} = 1$, two moderate effects of $\beta_{0j} = 0.5$, and six (for $p = 10$) or sixteen (for $p = 20$) noise parameters where $\beta_{0j} = 0$. To examine the effect of regularization we compare our CRs for the regularized estimators to CR^{OLS}, the normal CR based on the empirical standard error of the perturbed ordinary least squares (OLS) estimates.

Table 1.1: Coverage probabilities (lengths) of confidence regions when $\sigma = 1$. We multiply values by 100. The lengths of the simultaneous confidence regions are averaged over the number of parameters.

p	β_0		$n = \mathbf{100}$			$n = \mathbf{200}$			$n = \mathbf{400}$		
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
10	1	CR^{*N}	91.6 (38)	92.7 (41)	92.5 (52)	92.9 (27)	94.1 (29)	94.2 (37)	93.9 (19)	95.1 (21)	94.9 (26)
		CR^{*HDR}	91.4 (38)	91.7 (40)	91.5 (51)	92.4 (27)	93.7 (29)	93.9 (36)	93.4 (19)	94.5 (21)	94.1 (26)
		CR^{*Q}	91.7 (38)	91.0 (40)	91.5 (51)	92.5 (27)	93.6 (29)	93.8 (36)	93.5 (19)	94.3 (20)	93.9 (26)
		CR^{Asym}	93.9 (41)	94.1 (43)	93.7 (53)	94.1 (28)	94.2 (30)	94.1 (36)	94.4 (20)	95.0 (21)	94.9 (25)
		CR^{OLS}	91.4 (38)	91.7 (40)	90.6 (51)	93.0 (27)	93.6 (29)	93.9 (36)	93.0 (19)	94.3 (21)	93.7 (26)
	0.5	CR^{*N}	93.0 (40)	93.3 (43)	92.0 (54)	93.9 (28)	94.6 (30)	95.1 (38)	93.7 (20)	94.2 (21)	95.2 (27)
		CR^{*HDR}	91.9 (38)	92.5 (41)	93.4 (51)	93.5 (27)	94.0 (29)	93.8 (37)	93.7 (19)	93.8 (21)	94.7 (26)
		CR^{*Q}	91.7 (39)	92.3 (42)	90.7 (53)	93.3 (27)	93.7 (29)	93.9 (37)	93.8 (19)	93.6 (21)	94.7 (26)
		CR^{Asym}	93.3 (41)	93.5 (43)	91.5 (52)	94.5 (28)	94.3 (30)	93.7 (36)	95.0 (20)	94.0 (21)	94.1 (25)
		CR^{OLS}	92.4 (38)	91.6 (41)	90.9 (50)	93.5 (27)	94.3 (29)	93.8 (36)	94.3 (19)	93.7 (21)	94.9 (26)
	0	CR^{*N}	97.6 (23)	98.5 (25)	98.1 (31)	98.4 (17)	98.3 (19)	97.9 (23)	98.7 (13)	98.7 (13)	98.7 (16)
		CR^{*HDR}	99.1 (17)	99.3 (18)	99.4 (23)	99.6 (12)	99.3 (13)	99.0 (16)	99.7 (8)	99.3 (8)	99.5 (11)
		CR^{*Q}	99.5 (31)	99.7 (33)	99.8 (43)	99.7 (22)	99.7 (24)	99.7 (30)	99.8 (16)	99.7 (17)	99.8 (21)
		CR^{OLS}	92.9 (38)	92.9 (40)	91.7 (51)	93.4 (27)	93.0 (29)	92.6 (36)	93.4 (19)	93.7 (21)	93.6 (26)
		$CR^{*SimHDR}$	91.9 (36)	92.5 (39)	91.7 (49)	93.1 (26)	94.9 (28)	93.8 (36)	94.9 (19)	94.8 (20)	96.0 (25)
		CR^{*Sim}	92.5 (42)	92.9 (46)	91.9 (58)	93.7 (31)	95.5 (33)	95.2 (42)	95.5 (23)	95.7 (24)	96.6 (30)
		$CR^{*SimOLS}$	87.1 (54)	86.5 (58)	85.9 (72)	89.8 (38)	90.0 (41)	90.9 (52)	92.3 (28)	91.6 (30)	92.6 (37)
20	1	CR^{*N}	91.7 (38)	92.4 (42)	92.5 (53)	93.3 (27)	92.9 (30)	94.3 (38)	95.4 (19)	94.6 (21)	94.5 (27)
		CR^{*HDR}	90.2 (37)	90.9 (41)	90.8 (51)	92.4 (26)	91.9 (29)	91.9 (36)	95.1 (19)	93.6 (21)	93.3 (26)
		CR^{*Q}	90.2 (37)	90.7 (41)	90.7 (51)	92.3 (26)	92.3 (29)	91.9 (36)	95.1 (19)	93.2 (21)	93.1 (26)
		CR^{Asym}	93.9 (41)	94.5 (44)	93.3 (54)	95.1 (28)	93.9 (30)	94.0 (37)	95.9 (20)	94.7 (21)	93.3 (26)
		CR^{OLS}	90.3 (38)	89.9 (42)	90.0 (52)	92.6 (27)	92.1 (29)	92.0 (37)	95.3 (19)	93.4 (21)	93.3 (26)
	0.5	CR^{*N}	91.1 (40)	91.5 (44)	91.0 (56)	93.7 (28)	93.5 (30)	92.7 (39)	93.1 (20)	94.7 (21)	95.0 (27)
		CR^{*HDR}	89.7 (38)	90.3 (42)	92.5 (52)	93.1 (27)	93.1 (30)	91.7 (38)	92.9 (19)	93.3 (21)	94.4 (27)
		CR^{*Q}	89.7 (39)	89.7 (43)	89.2 (54)	92.7 (27)	92.5 (30)	91.5 (38)	92.7 (19)	93.5 (21)	94.5 (27)
		CR^{Asym}	91.7 (41)	92.0 (44)	89.7 (53)	94.5 (28)	93.2 (30)	92.3 (37)	93.7 (20)	94.3 (21)	94.2 (26)
		CR^{OLS}	89.7 (38)	89.7 (42)	89.6 (52)	92.9 (27)	92.9 (29)	91.8 (37)	92.9 (19)	92.7 (21)	94.2 (26)
	0	CR^{*N}	96.6 (29)	97.3 (32)	96.8 (40)	98.6 (21)	98.5 (23)	98.7 (29)	98.8 (15)	99.1 (17)	99.0 (21)
		CR^{*HDR}	97.7 (25)	98.3 (28)	98.3 (34)	98.9 (17)	99.3 (19)	99.1 (24)	99.3 (12)	99.5 (13)	99.4 (16)
		CR^{*Q}	99.0 (31)	99.4 (34)	99.2 (43)	99.5 (22)	99.9 (24)	99.5 (30)	99.8 (15)	99.9 (17)	99.7 (21)
		CR^{OLS}	89.5 (38)	90.1 (41)	90.1 (52)	92.4 (27)	92.2 (29)	92.6 (37)	94.7 (19)	93.8 (21)	94.2 (26)
		$CR^{*SimHDR}$	90.5 (42)	92.0 (47)	92.1 (58)	95.4 (31)	95.7 (34)	95.3 (44)	96.9 (22)	96.9 (25)	96.9 (32)
		CR^{*Sim}	92.5 (51)	91.9 (57)	91.7 (71)	96.5 (38)	97.5 (42)	96.5 (54)	97.7 (28)	97.8 (31)	98.1 (40)
		$CR^{*SimOLS}$	80.1 (59)	78.4 (64)	77.9 (80)	87.5 (41)	87.1 (45)	84.9 (57)	90.6 (29)	90.0 (32)	91.1 (41)

Table 1.2: Coverage probabilities (lengths) of confidence regions when $\sigma = 2$. We multiply values by 100. The lengths of the simultaneous confidence regions are averaged over the number of parameters.

p	β_0		$n = \mathbf{100}$			$n = \mathbf{200}$			$n = \mathbf{400}$		
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
10	1	CR^{*N}	92.6 (79)	94.3 (85)	92.9 (110)	93.4 (55)	94.4 (59)	94.9 (76)	94.4 (39)	94.2 (42)	94.7 (53)
		CR^{*HDR}	91.7 (76)	93.9 (82)	94.0 (104)	92.7 (55)	94.2 (59)	94.4 (74)	94.0 (39)	93.6 (42)	93.7 (52)
		CR^{*Q}	91.9 (77)	93.7 (84)	91.4 (107)	92.7 (55)	93.8 (59)	94.3 (75)	93.9 (39)	93.4 (42)	93.8 (52)
		CR^{Asym}	80.7 (57)	82.7 (60)	78.8 (73)	82.3 (40)	83.6 (42)	80.5 (51)	83.7 (28)	83.5 (30)	80.9 (36)
		CR^{OLS}	91.7 (75)	93.4 (81)	91.1 (102)	92.7 (54)	93.4 (58)	94.0 (73)	94.0 (39)	93.7 (42)	93.4 (52)
	0.5	CR^{*N}	87.5 (80)	87.6 (86)	81.3 (100)	93.5 (59)	94.2 (63)	90.3 (79)	94.7 (41)	95.7 (44)	94.5 (57)
		CR^{*HDR}	90.3 (71)	91.4 (76)	83.9 (88)	95.9 (54)	96.5 (58)	92.3 (69)	93.5 (39)	94.5 (42)	96.5 (52)
		CR^{*Q}	91.5 (76)	92.3 (81)	90.5 (97)	93.4 (57)	93.9 (61)	92.8 (75)	94.0 (40)	94.6 (43)	94.0 (56)
		CR^{Asym}	76.5 (51)	76.3 (53)	67.3 (57)	78.8 (39)	78.8 (42)	78.3 (48)	79.9 (28)	80.9 (29)	78.1 (36)
		CR^{OLS}	91.1 (75)	92.1 (81)	90.7 (101)	93.0 (54)	93.5 (58)	93.1 (72)	94.1 (38)	94.5 (41)	93.5 (52)
	0	CR^{*N}	97.1 (47)	97.1 (50)	97.7 (65)	98.1 (33)	98.0 (37)	98.1 (47)	98.3 (25)	98.4 (27)	98.8 (34)
		CR^{*HDR}	98.3 (40)	98.2 (40)	99.1 (52)	99.0 (25)	99.4 (28)	99.5 (36)	99.4 (17)	99.5 (19)	99.6 (24)
		CR^{*Q}	99.1 (64)	98.9 (68)	99.4 (86)	99.7 (45)	99.7 (49)	99.7 (61)	99.9 (32)	99.7 (34)	99.9 (43)
		CR^{OLS}	91.3 (75)	91.2 (81)	91.7 (102)	92.9 (54)	92.5 (58)	92.3 (73)	94.1 (39)	94.7 (42)	94.2 (52)
		$CR^{*SimHDR}$	85.3 (71)	85.7 (77)	75.1 (96)	94.1 (51)	94.4 (56)	90.8 (70)	96.3 (38)	95.2 (41)	95.2 (52)
		CR^{*Sim}	84.1 (83)	84.1 (91)	73.9 (116)	93.4 (60)	93.5 (66)	90.1 (84)	96.6 (45)	95.9 (49)	95.8 (62)
		$CR^{*SimOLS}$	85.2 (108)	86.9 (116)	87.5 (146)	90.9 (77)	90.7 (83)	91.1 (104)	92.7 (55)	92.5 (59)	92.6 (74)
20	1	CR^{*N}	91.2 (80)	91.7 (87)	90.0 (112)	92.1 (55)	93.9 (60)	92.9 (78)	94.3 (39)	93.7 (43)	94.6 (54)
		CR^{*HDR}	90.2 (76)	91.2 (83)	92.1 (104)	92.1 (54)	93.1 (59)	91.7 (75)	94.1 (38)	93.1 (42)	93.9 (53)
		CR^{*Q}	90.1 (77)	90.7 (84)	89.6 (107)	92.1 (54)	93.3 (59)	91.9 (76)	94.4 (38)	92.6 (42)	93.8 (53)
		CR^{Asym}	79.2 (58)	78.9 (62)	75.9 (76)	80.3 (40)	82.2 (43)	77.3 (53)	83.5 (28)	80.5 (30)	81.9 (37)
		CR^{OLS}	89.7 (76)	90.1 (83)	89.7 (104)	92.3 (54)	92.9 (58)	91.3 (74)	94.4 (38)	92.9 (42)	93.8 (53)
	0.5	CR^{*N}	86.1 (81)	82.7 (86)	80.7 (103)	91.6 (59)	91.9 (64)	88.4 (81)	93.9 (41)	94.9 (45)	93.1 (58)
		CR^{*HDR}	91.0 (74)	87.8 (79)	84.7 (94)	95.7 (55)	94.7 (59)	92.1 (73)	93.3 (39)	94.3 (43)	96.1 (54)
		CR^{*Q}	89.1 (75)	88.8 (80)	88.6 (97)	92.0 (57)	92.1 (61)	90.9 (75)	93.3 (40)	94.1 (44)	92.5 (56)
		CR^{Asym}	72.5 (50)	71.6 (51)	64.2 (57)	77.2 (39)	76.6 (41)	73.8 (48)	81.3 (28)	79.2 (30)	76.8 (36)
		CR^{OLS}	89.5 (76)	88.9 (82)	89.3 (104)	92.5 (54)	92.5 (59)	91.8 (74)	94.1 (38)	94.4 (42)	93.1 (53)
	0	CR^{*N}	97.3 (57)	96.8 (61)	97.0 (79)	98.5 (40)	97.3 (45)	97.7 (58)	98.8 (29)	98.9 (33)	98.9 (41)
		CR^{*HDR}	97.9 (53)	97.5 (55)	97.7 (70)	99.1 (35)	98.2 (39)	98.7 (50)	99.3 (24)	99.3 (27)	99.3 (33)
		CR^{*Q}	98.9 (63)	98.7 (69)	98.8 (87)	99.6 (44)	99.2 (49)	99.3 (62)	99.6 (31)	99.7 (34)	99.7 (43)
		CR^{OLS}	90.1 (76)	90.2 (83)	89.8 (104)	92.8 (54)	91.9 (59)	91.9 (74)	93.5 (38)	93.5 (42)	93.7 (53)
		$CR^{*SimHDR}$	87.1 (81)	83.6 (89)	78.1 (112)	95.1 (60)	94.4 (66)	93.7 (84)	97.1 (45)	97.1 (50)	97.3 (62)
		CR^{*Sim}	86.1 (99)	82.8 (109)	78.3 (138)	94.3 (73)	94.4 (81)	94.1 (103)	97.8 (55)	98.3 (61)	97.1 (77)
		$CR^{*SimOLS}$	77.7 (117)	76.3 (128)	76.6 (160)	86.1 (82)	85.4 (90)	86.9 (114)	90.5 (59)	89.8 (64)	90.3 (81)

In Tables 1.1 and 1.2 we see that when $\sigma = 1$, most regions perform similarly for nonzero parameters. When $\sigma = 2$, the perturbation regions usually have higher coverage than CR^{Asym} and sacrifice very little in length. The asymmetric $\text{CR}^{*\text{HDR}}$ has the shortest length when $\beta_{0j} = 0$ for all settings. Coverage for $\text{CR}^{*\text{HDR}}$ and simultaneous confidence regions can be low when $n = 100$ due to the difficulty of estimating $\widehat{\mathcal{P}}_{0j}$ at such a small sample size, but coverage reaches nominal levels by $n = 200$. The standard deviation estimate from Zou (2006), $\widehat{\sigma}^{\text{Asym}}$ (also see Table 1.3), is not large enough to cover β_{0j} sufficiently, and while the coverage probability of the CR^{OLS} is not extremely low, it is notably outperformed by the other confidence regions when $\beta_{0j} = 0$. We omit the results from the settings where $n = 1000$ because the results have similar patterns as those with $n = 400$. For these large sample cases with n greater than or equal to 400 we saw convergence to 95% coverage for the normal CRs, highest density regions, and OLS CRs in all settings when the true parameter was nonzero. For true zero parameters, the coverage probabilities of our confidence regions converged to 1, while the OLS CR converged to 0.95. A tradeoff associated with our method is that while the coverage of our perturbation confidence regions tends to be higher than CR^{OLS} and CR^{Asym}, some power is sacrificed for moderate signals of $\beta_{0j} = 0.5$. This loss is minimal, however, and only appears in difficult cases when sample size is low and ρ and σ are high. When $\beta_{0j} = 0$, CR^{OLS} has coverage lower than 95% for small samples while our methods produce regions with coverage probability near 1 and very short lengths reflecting the oracle properties. Overall, the most disparity between our methods and previous methods is seen when the SNR is low.

The coverage probabilities and lengths of our simultaneous confidence regions are also displayed in Tables 1.1 and 1.2. We compared our methods to $\text{CR}^{*\text{SimOLS}}$,

constructed analogously to $\text{CR}^{*\text{Sim}}$ except $\widehat{\mathcal{A}}^* = \{j|j = 1, \ldots, p\}$ and $\text{CR}^{*\text{SimOLS}}$ is centered at the OLS estimates and the standard error is the sample standard deviation of the perturbed OLS estimates. Our regularized $\text{CR}^{*\text{Sim}}$ and $\text{CR}^{*\text{SimHDR}}$ have the advantage of shrinking the dimension of the region by reducing some CRs to the point $\{0\}$ when $\widehat{\mathcal{P}}_{0j}$ is large. We see that our $\text{CR}^{*\text{Sim}}$ and $\text{CR}^{*\text{SimHDR}}$ outperform $\text{CR}^{*\text{SimOLS}}$ in coverage and have shorter lengths. For large sample settings when $n = 1000$, $\text{CR}^{*\text{SimOLS}}$ converges further to 95% coverage with levels around 90% for $p = 20$ and $\text{CR}^{*\text{Sim}}$ and $\text{CR}^{*\text{SimHDR}}$ have coverage almost always over 95%.

In Table 1.3 we also present the standard error estimates when $\sigma = 2$. For notation, let the empirical standard deviations of the estimators $\widehat{\beta}_j$ and $\widetilde{\beta}_j$ be denoted as $\widetilde{\sigma}_j$ and $\widetilde{\sigma}_j^{\text{OLS}}$, respectively. We see that our estimate of the standard error from the perturbed samples, $\widehat{\sigma}_j^*$, does well in estimating $\widetilde{\sigma}_j$. However, the standard error proposed by Zou (2006) underestimates the true standard error of the parameter estimates, especially when $\sigma = 2$ and $\beta_{0j} = 0.5$ or 0. When the SNR is higher, $\widetilde{\sigma}_j^{\text{Asym}}$ estimates $\widetilde{\sigma}_j$ well except when $\beta_{0j} = 0$ because $\hat{\sigma}_j^{\text{Asym}} = 0$ whereas $\widetilde{\sigma}_j$ and $\widehat{\sigma}_j^*$ are clearly nonzero.

1.4 Example: HIV drug resistance

We illustrate our methods in a real example using the HIV antiretroviral drug susceptibility data described in Rhee et al. (2003). This dataset was refined from the Stanford HIV Drug Resistance Database (available at *http://hivdb.stanford.edu/*), and is used to study the association of protease mutations with susceptibility to the protease inhibitor anti-retroviral (ARV) drug amprenavir. The data consist of mutation information at 99 protease codons in the viral genome, of which 79 contain

Table 1.3: Empirical s.d. of the parameter estimates ($\widetilde{\sigma}$) and average s.e. estimates ($\hat{\sigma}$). We present results for settings when σ, the standard deviation of ϵ, is 2. All values are multiplied by 100. Note that $\hat{\sigma}_j^{\text{Asym}} = 0$ when $\hat{\beta}_j = 0$, but $\hat{\beta}_j$ and $\hat{\beta}_j^*$ are not always 0 in the simulations, and therefore the average $\hat{\sigma}_j^{\text{Asym}}$ is nonzero.

p	β_0		$n = \mathbf{100}$			$n = \mathbf{200}$			$n = \mathbf{400}$		
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
10	1	$\widetilde{\boldsymbol{\sigma}}$	21.7	22.2	29.9	14.7	15.2	19.3	10.1	10.8	13.7
		$\widetilde{\boldsymbol{\sigma}}^{\text{OLS}}$	21.1	21.6	28.5	14.4	15.2	19.1	10.1	10.9	13.9
		$\widehat{\boldsymbol{\sigma}}^*$	20.0	21.7	28.1	14.1	15.2	19.4	10.0	10.8	13.5
		$\widehat{\boldsymbol{\sigma}}^{\text{OLS}}$	19.1	20.6	26.0	13.8	14.8	18.5	9.8	10.6	13.2
		$\widehat{\boldsymbol{\sigma}}^{\text{Asym}}$	14.7	15.4	18.7	10.2	10.8	13.1	7.1	7.5	9.2
	0.5	$\widetilde{\boldsymbol{\sigma}}$	24.2	25.5	32.3	16.1	16.8	21.8	10.6	11.2	14.7
		$\widetilde{\boldsymbol{\sigma}}^{\text{OLS}}$	21.6	22.8	29.2	14.8	15.5	19.4	10.2	10.8	13.9
		$\widehat{\boldsymbol{\sigma}}^*$	21.1	22.8	27.9	15.1	16.2	20.5	10.3	11.2	14.5
		$\widehat{\boldsymbol{\sigma}}^{\text{OLS}}$	19.1	20.7	25.9	13.8	14.8	18.4	9.8	10.6	13.2
		$\widehat{\boldsymbol{\sigma}}^{\text{Asym}}$	13.0	13.6	14.5	10.0	10.6	12.1	7.1	7.5	9.1
	0	$\widetilde{\boldsymbol{\sigma}}$	17.0	17.2	20.3	10.7	11.5	14.2	6.9	7.2	9.2
		$\widetilde{\boldsymbol{\sigma}}^{\text{OLS}}$	22.4	23.5	28.4	14.9	16.2	19.9	10.2	10.9	13.8
		$\widehat{\boldsymbol{\sigma}}^*$	18.6	19.9	25.1	13.2	14.2	17.8	9.4	10.1	12.6
		$\widehat{\boldsymbol{\sigma}}^{\text{OLS}}$	19.2	20.6	26.1	13.8	14.9	18.5	9.9	10.6	13.3
		$\widehat{\boldsymbol{\sigma}}^{\text{Asym}}$	5.0	4.5	5.4	2.6	2.9	3.5	1.5	1.5	2.0
20	1	$\widetilde{\boldsymbol{\sigma}}$	23.2	24.8	33.4	15.4	16.2	21.6	9.9	11.4	14.0
		$\widetilde{\boldsymbol{\sigma}}^{\text{OLS}}$	23.0	24.6	31.7	15.4	16.1	21.3	9.8	11.4	13.9
		$\widehat{\boldsymbol{\sigma}}^*$	20.3	22.2	28.5	14.1	15.3	19.9	9.9	10.9	13.9
		$\widehat{\boldsymbol{\sigma}}^{\text{OLS}}$	19.3	21.1	26.5	13.7	14.9	18.9	9.7	10.7	13.4
		$\widehat{\boldsymbol{\sigma}}^{\text{Asym}}$	14.9	15.9	19.4	10.3	10.9	13.5	7.1	7.6	9.4
	0.5	$\widetilde{\boldsymbol{\sigma}}$	24.7	27.1	32.8	16.4	18.1	23.1	10.5	11.7	15.7
		$\widetilde{\boldsymbol{\sigma}}^{\text{OLS}}$	22.8	25.3	31.6	15.1	16.6	20.8	10.0	11.2	14.4
		$\widehat{\boldsymbol{\sigma}}^*$	21.2	22.8	28.1	15.2	16.5	20.8	10.5	11.5	14.9
		$\widehat{\boldsymbol{\sigma}}^{\text{OLS}}$	19.3	21.0	26.4	13.7	14.9	18.8	9.8	10.7	13.5
		$\widehat{\boldsymbol{\sigma}}^{\text{Asym}}$	12.8	13.1	14.5	10.1	10.5	12.2	7.2	7.6	9.2
	0	$\widetilde{\boldsymbol{\sigma}}$	15.3	16.7	21.1	9.3	10.8	13.7	6.1	6.6	8.1
		$\widetilde{\boldsymbol{\sigma}}^{\text{OLS}}$	22.5	25.0	31.7	14.7	16.6	21.2	10.2	11.4	14.1
		$\widehat{\boldsymbol{\sigma}}^*$	18.5	20.2	25.6	12.9	14.2	18.1	9.1	10.2	12.7
		$\widehat{\boldsymbol{\sigma}}^{\text{Asym}}$	4.7	5.1	6.1	2.6	2.8	3.5	1.4	1.6	1.9

mutations, and ARV drug resistance assays for $n = 702$ HIV infected patients. Drug resistance was measured in units of IC_{50}, the amount of drug needed to inhibit viral replication by 50% in units of fold increase compared to drug-sensitive wildtype virus. Researchers are interested in determining which protease mutations are associated with ARV resistance so that they may develop a genotype test for resistance that looks for these mutations in the patient's infecting HIV strain. Therefore, we aim to examine the effect of the presence of any of the mutations at 79 codons on IC_{50}, where higher IC_{50} measurements indicate higher levels of drug resistance. We chose to log-transform the non-negative IC_{50} outcome and represented the presence of each of the mutations as a binary predictor in our regression model. We removed the fifteen mutations that occurred less than 0.5% in the data set. Recently, Wu (2009) analyzed these data with a permutation test for regression coefficients of LASSO. In this paper, we will analyze the data using ALASSO and gain inference by using our perturbation methods to construct CRs and standard errors.

For this analysis, we used LARS to fit an ALASSO linear model with initial parameters $\widetilde{\boldsymbol{\beta}}$ estimated by OLS and λ and λ^* chosen to minimize the BIC as described in the appendix. We generated M=500 perturbation variable sets $\mathcal{G}$, consisting of $n = 702$ *i.i.d.* variables from an exponential distribution with mean and variance equal to 1, and for each $\mathcal{G}$ we minimized the perturbed objective function to obtain $\widehat{\boldsymbol{\beta}}^*_m$. We constructed 95% CRs using our perturbation method and compared inference gained from CR^{*N}, CR^{*HDR}, and CR^{*Q} to the inference from CR^{Asym} and CR^{OLS}. We estimated the σ used in the standard deviation estimate from Zou (2006) analogous to equation (1.2) with the residual variance from the nonregularized linear regression model and chose $\widehat{p}_{high}$ and $\widehat{p}_{low}$ as described in the simulation studies.

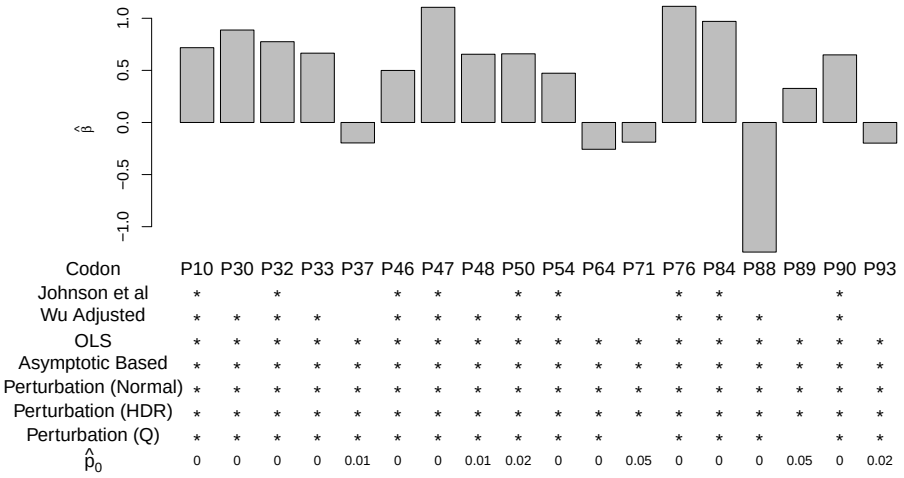

Figure 1.1: Perturbation methods results denoting significant associations between genetic mutations and drug susceptibility.

We present a graphical summary of the analysis results in Figure 1.4. Previous studies by Prado et al. (2002) and results collected by Johnson et al. (2005) found that mutations at codons 10, 32, 46, 47, 50, 54, 73, 82, 84 and 90 emerge in amprenavir resistant viral genomes. Using a permutation based p-value adjusted for multiple testing, Wu (2009) determined these mutations (except 73 and 82) as well as additional codon mutations to be significantly associated with amprenavir susceptibility at the $\alpha = 0.05$ level for a total of thirteen significant associations. The ALASSO estimator obtained with $\lambda = 0.56$ from BIC estimated 36 coefficients as nonzero. The confidence region from nonregularized estimates $\mathrm{CR}^{\mathrm{OLS}}$ was significant for twenty-six mutations. Our perturbation based $\mathrm{CR}^{*\mathrm{N}}$, $\mathrm{CR}^{*\mathrm{HDR}}$, and $\mathrm{CR}^{*\mathrm{Q}}$ for the mutations found significant by Wu (2009) did not include zero and three new mutations (37, 64, 93) had significant perturbation confidence regions. We see in Figure 1.4 that the parameter for codons 71 and 89 have marginally significant Normal and

HDR confidence regions and marginally nonsignificant quantile confidence regions and note that $\widehat{\mathcal{P}}_{0j}$ is marginally close to 0.05.

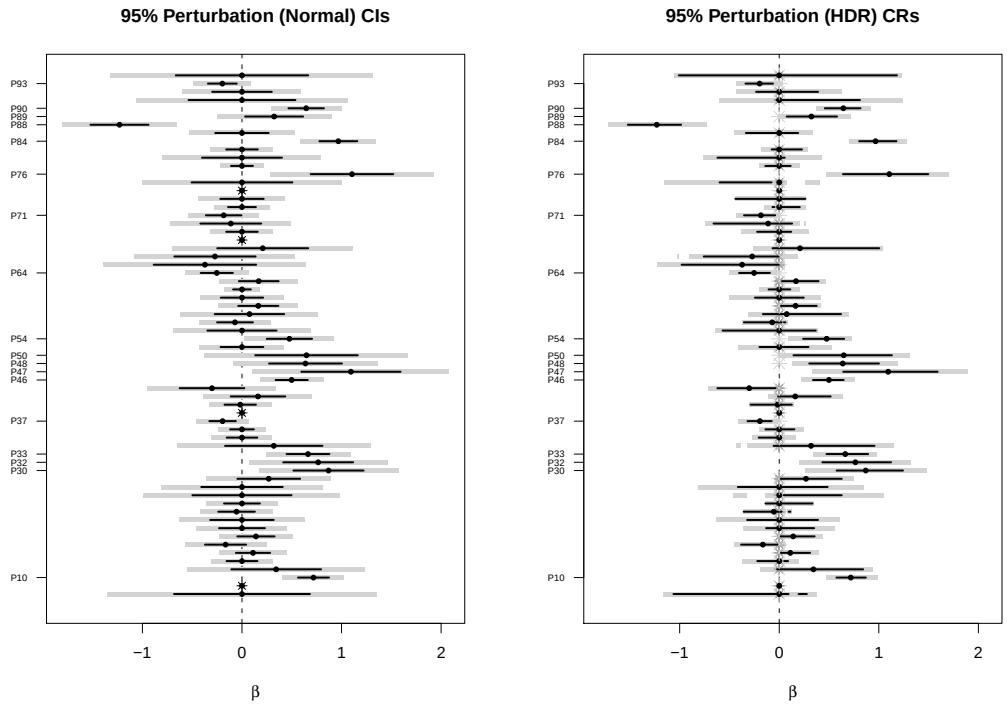

Figure 1.2: 95% perturbation CRs ($\text{CR}^{*\text{N}}$ and $\text{CR}^{*\text{HDR}}$) for the association between genetic mutations and antiretroviral drug susceptibility. Estimated coefficients $\widehat{\beta}_j$ are represented with a circle on each CR line and a star at zero signifies that the CR includes the point mass at zero. The shaded region denotes the simultaneous confidence regions $\text{CR}^{*\text{Sim}}$ and $\text{CR}^{*\text{SimHDR}}$. Note that even coefficients estimated as zero may have CRs around their estimates and that $\text{CR}^{*\text{HDR}}$ may be asymmetrical and noncontiguous.

Our use of ALASSO provides estimates of the effects of each mutation while adjusting for the presence of other mutations. Several studies have shown that mutations associated with resistance to protease inhibitors can have varying effects when combined with other mutations (Schumi and DeGruttola, 2008; Van Marck et al., 2009). For instance, the mutation at codon 32 has been found to have no effect on

resistance of the protease inhibitor drug darunavir when a mutation at codon 84 is present (Van Marck et al., 2009). Our method allows us to determine the size of associations without orthogonalizing predictors and we adjust for multiple testing with the simultaneous confidence region CR^{*Sim}. Results could be impacted by studies summarized in Johnson et al. (2005) that may not have adjusted for other mutations, and the use of LASSO estimators that do not have oracle properties in Wu (2009). Our methods highlight three new mutations that have not been found to be associated with drug susceptibility. Furthermore, our methods produce CRs for the coefficients of mutations that were estimated as zero. These CRs quantify the uncertainty in our estimation and can aid scientists who wish to conduct future drug therapy studies involving the codons.

1.5 Discussion

In this paper, we address the problem of constructing a covariance estimate for parameter estimates obtained with a general objective function and concave penalty functions including adaptive LASSO and SCAD. The proposed methods for covariance estimates are simple to implement and possess the attractive property that parameters estimated as zero have nonzero standard errors. We may then construct confidence regions for each parameter estimate and obtain more meaningful inference.

We have shown through extensive simulation studies using the ALASSO penalty that our perturbation method results in confidence regions with accurate coverage probability. The perturbation based normal CR does not sacrifice much in length and has reasonable coverage for small sample sizes. We set the CR to $\{0\}$ when the proportion of perturbed estimates set to 0 is higher than a threshold, and therefore

shorten the length by utilizing the oracle property. The perturbation based highest density region has even shorter length and good coverage probability, especially for the moderate signal $\beta_{0j} = 0.5$ in comparison to all other confidence regions. The asymptotic based Normal interval that uses the standard error estimate presented in Zou (2006) fails to reach nominal coverage levels due to the underestimation of the standard error, most notably when the standard error is estimated as 0 when $\widehat{\boldsymbol{\beta}} = 0$. However, our estimate of the standard error of the parameter estimates based on our perturbation samples is close to the empirical standard error of the ALASSO estimates, even for parameters estimated as 0. Additionally, we propose two types of simultaneous CRs that adjust for multiple comparisons. We again utilize the oracle property and reduce the dimension of our region by setting intervals to $\{0\}$ when the proportion of zero perturbed parameter estimates is high. Therefore, the average length of our Normal simultaneous region will often be shorter than the simultaneous OLS region. For instance, when all covariates are independent, the OLS length is asymptotically proportional to $\gamma_{\text{OLS}} = \max\left\{\left|(\widetilde{\beta}_j - \beta_{0j})/\sigma\right|\right\}_{j=1}^{p}$ whereas the perturbation region length is asymptotically proportional to $(q/p)\gamma$ where $\gamma = \max\left\{\left|(\widehat{\beta}_j - \beta_{0j})/\sigma\right|\right\}_{\beta_{0j}\neq 0}$. Note that $\gamma \leq \gamma_{\text{OLS}}$ and so the length of the perturbation region will be shorter than the OLS length when the true model is sparse. Similarly, when the covariates are not independent, $\left\{(\widetilde{\beta}_j - \beta_{0j})/\sigma\right\}_{j=1}^{p} \sim \mathcal{N}(\mathbf{0}, \text{Corr}(\widehat{\boldsymbol{\beta}}))$ and the perturbation region generally has shorter average length than the OLS region. Simple simulations show that when q parameters are estimated as nonzero, we expect the perturbation region length to be approximately 0.36 times the OLS region length when $p = 10$ and $q = 4$ and approximately 0.16 times the OLS region length when $p = 20$ and $q = 4$ for both the independent case and the compound symmetry case when $\rho = 0.5$ and $\sigma = 1$. However, in finite sample, the gain in interval length for

the shrinkage estimators may be substantially less than the theoretical gain as oracle properties may be far from being true and the intervals may need to be enlarged to ensure proper coverage levels.

When the SNR is low, much larger sample sizes may be required for the re-sampling procedure to yield confidence intervals with proper coverage levels. We conducted further simulations for the case when $\boldsymbol{\beta} = (1, 1, 0.1, 0.1, \mathbf{0}_{1\times(p-4)})^{\mathsf{T}}$. In general, we find that the standard error estimates perform well even with sample sizes around 100. The confidence intervals have reasonable coverage levels for β_3 when $\sigma = 1$ and the correlation ρ is low with sample size 400 or larger. For example, when $\sigma = 1$, $\rho = 0.2$, and $p = 20$, the coverage level of the 95% $\mathrm{CR}^{*\mathrm{HDR}}$ of β_3 is about 90% for $n = 400$ and 94% for $n = 1000$. As we increase the correlation ρ and σ, the interval estimation of β_3 becomes more difficult. For example, for the most difficult case with $\sigma = 2$, $\rho = 0.5$, and $p = 20$, the empirical coverage level of $\mathrm{CR}^{*\mathrm{HDR}}$ is about 60%, 84% and 90% when $n = 400$, 1000, and 2000, respectively. This is a particularly difficult case as it has been shown that estimating the distribution function of the ALASSO type estimator is not feasible when the effect size is of similar magnitude to $n^{-\frac{1}{2}}$ (Pötscher and Schneider, 2009). Note that when $\sigma = 2$, the effect size corresponding to β_3 is 0.05 whereas $n^{-\frac{1}{2}} = 0.1$ when n is 100 and $n^{-\frac{1}{2}} \approx 0.032$ for $n = 1000$.

Additionally, it is well known that regularized estimators, while possessing asymptotic oracle properties, are prone to bias in finite samples. Bias correction for the ALASSO estimator can be achieved based on our perturbation samples. We present the technical details of the estimation of the bias in the appendix. We find that this bias correction works well in practice, especially when the signal

is small or moderate, as when $\beta_{0j} = 0.5$. For example, in our simulations when $p = 20, n = 200, \rho = 0.2, \sigma = 2$, and $\beta_{0j} = 0.5$, the bias of $\widehat{\beta}_j$ is -0.067 while the bias of $\widehat{\beta}_{0j}^{\text{BC}}$ is -0.034. Similar gains are seen for most settings. The bias corrected estimator has empirical standard error similar to that of the original ALASSO estimator but with smaller bias. We could construct analogous bias-corrected estimators based on other penalties and objective functions. The model size with the ALASSO and bias-corrected ALASSO estimator in our simulations is close to 5 when $\sigma = 1$, except for the difficult cases when $n = 100$ and $p = 20$ for which the average model size is closer to 5.5. For the settings where the SNR is low with $\sigma = 2$, the oracle property is weak in finite samples and so the model size is between 5 and 6 when $p = 10$ and between 6 and 9 when $p = 20$.

We note that when p is large relative to n, initial parameter estimates obtained with ridge regression can produce more stable results. Furthermore, our results may be extended to the case where p tends to ∞ at some rate slower than n. We expect that the theory could be derived using similar arguments as given in Fan and Peng (2004) and Zou and Zhang (2009). Lastly, we note that our methods are robust to misspecification of the model and are valid provided that regularity conditions given in Section 2 hold.

1.6 Acknowledgement

The authors thank the editor, the associate editor, and two referees for their insightful and constructive comments that greatly improved the article.

1.7 Appendix A: Proofs

1.7.1 Justification for the resampling method

To show that the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ can be estimated by that of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*-\widehat{\boldsymbol{\theta}}) \mid \mathcal{X}$ under conditions C1-C3, we first consider the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0)$ under the product probability measure $\mathbb{P}^*$ generated by the data, $\mathcal{X}$, and the perturbation variables $\mathcal{G} = \{G_i, i = 1, \ldots, n\}$. Throughout, we assume that the parameter space for $\boldsymbol{\theta}$, denoted by Ω, is a compact set and $\boldsymbol{\theta}_0$ is an interior point of Ω. Note that this compactness condition may be nontrivial in practice. This condition is necessary for this proof of our proposed method, and validity of the method without this condition warrants further investigation. We let $\mathbb{P}_n$ denote the empirical measure generated by $\mathcal{X}$ and $\mathbb{G}_n = n^{-\frac{1}{2}}(\mathbb{P}_n - \mathbb{P})$. We use notation $\rightarrow_p$ to denote convergence in probability.

We first show that $\widetilde{\boldsymbol{\theta}}^* \rightarrow_p \boldsymbol{\theta}_0$, where $\widetilde{\boldsymbol{\theta}}^*$ is the perturbed initial parameter estimate obtained by minimizing the perturbed un-regularized likelihood in (1.3). For $f \in \{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}$, denote the empirical perturbation measure as $\mathbb{P}_n^* f = n^{-1}\sum_{i=1}^n G_i f(X_i)$. Since $\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) : \boldsymbol{\theta} \in \Omega\}$ is $\mathbb{P}$-Glivenko-Cantelli, by Corollary 10.14 of Kosorok (2008) $|\widetilde{\mathcal{L}}^*(\boldsymbol{\theta}) - \mathbb{P}\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}| \leq |\widetilde{\mathcal{L}}^*(\boldsymbol{\theta}) - \widetilde{\mathcal{L}}(\boldsymbol{\theta})| + |\widetilde{\mathcal{L}}(\boldsymbol{\theta}) - \mathbb{P}\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}| = |(\mathbb{P}_n^* - \mathbb{P}_n)\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}| + |(\mathbb{P}_n - \mathbb{P})\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}|$ uniformly converges to zero. Then, under condition C1, $\mathbb{P}\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}$ has a unique minimum at $\boldsymbol{\theta}_0$, and so $\widetilde{\boldsymbol{\theta}}^* \rightarrow_p \boldsymbol{\theta}_0$ (Newey and McFadden, 1994, Theorem 2.1).

We now show that $\widehat{\boldsymbol{\theta}}^* \rightarrow_p \boldsymbol{\theta}_0$. First note that $\sum_{j=1}^p p'_{\lambda_n^* j}(|\widetilde{\beta}_j^*|)|\beta_j| \rightarrow 0$ in probability. When the penalty is L_q, $p'_{\lambda_n^* j}(|\widetilde{\beta}_j^*|) = \lambda_n |\beta_j|^q$, $p'(|\widetilde{\beta}_j^*|) \rightarrow_p p'(|\beta_{0j}|)$ by the continuous mapping theorem and $\lambda_n \rightarrow 0$. For the SCAD penalty, $p'_{\lambda_n^* j}(|\widetilde{\beta}_j^*|) = \lambda_n I(|\widetilde{\beta}_j^*| \leq \lambda_n) + (a\lambda_n - |\widetilde{\beta}_j^*|)_+ I(|\widetilde{\beta}_j^*| > \lambda_n)/(a-1)$. We consider two cases: (i) $\beta_{0j} \neq 0$, and

(ii) $\beta_{0j} = 0$. For case (i), $\lambda_n \to 0$ and $|\widetilde{\beta}_j^*| \to_p |\beta_{0j}|$. Thus, $I(|\widetilde{\beta}_j^*| \leq \lambda_n) \to_p 0$ and $(a\lambda_n - |\widetilde{\beta}_j^*|)_+ \to_p 0$. For case (ii), $\lambda_n \to 0$ and $(a\lambda_n - |\widetilde{\beta}_j^*|)_+ \to_p 0$. Finally, for the ALASSO penalty, $p'_{\lambda_{n}^* j}(|\widetilde{\beta}_j^*|) = \lambda_n |n^{\frac{1}{2}}\widetilde{\beta}_j^*|^{-1}$, $|n^{\frac{1}{2}}\widetilde{\beta}_j^*| = O_{\mathbb{P}}(1)$, and $\lambda_n \to 0$. Then, since $\boldsymbol{\theta}$ lies in a compact space, $\sum_{j=1}^p p'_{\lambda_{n}^* j}(|\widetilde{\beta}_j^*|)|\beta_j| \leq \tau \sum_{j=1}^p p'_{\lambda_{n}^* j}(|\widetilde{\beta}_j^*|) \leq ||\boldsymbol{\beta}||B_n$, where $\tau = \max\{|\beta_j|\}$, $B_n = o_{\mathbb{P}}(1)$ since $p'_{\lambda_{n}^* j}(|\widetilde{\beta}_j^*|) \to_{\mathbb{P}} 0$ for each j, and hence $\sup_{\boldsymbol{\theta}} \left| \sum_{j=1}^p p'_{\lambda_{n}^* j}(|\widetilde{\beta}_j^*|)|\beta_j| \right| \to_p 0$ (Newey and McFadden, 1994, Lemma 2.9). Now, with similar arguments as above for the proof of $\widetilde{\boldsymbol{\theta}}^* \to_p \boldsymbol{\theta}_0$, we have that $|\widehat{\mathcal{L}}^*(\boldsymbol{\theta}) - \mathbb{P}\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}| \leq |\widetilde{\mathcal{L}}^*(\boldsymbol{\theta}) - \mathbb{P}\{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})\}| + \sum_{j=1}^p p'_{\lambda_{n}^* j}(|\widetilde{\beta}_j^*|)|\beta_j|$ uniformly converges to zero. This implies the convergence of $\widehat{\boldsymbol{\theta}}^* \to_p \boldsymbol{\theta}_0$.

We next show that $\|\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0\| = O_{\mathbb{P}^*}(n^{-\frac{1}{2}})$. It is sufficient to show that for any $\epsilon > 0$, there exits $C > 0$ such that

$$\mathbb{P}^* \left(\inf_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq C n^{-\frac{1}{2}}} \widehat{\mathcal{L}}^*(\boldsymbol{\theta}) > \widehat{\mathcal{L}}^*(\boldsymbol{\theta}_0) \right) > 1 - \epsilon \tag{1.6}$$

Consider $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}$. Condition C3(c) implies

$$\frac{\mathbb{P}_n\{\mathcal{L}(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}) - \mathcal{L}(\boldsymbol{\theta}_0) - n^{-\frac{1}{2}}\mathcal{U}(\boldsymbol{\theta}_0; \mathcal{D})^\intercal \mathbf{u}\} - \frac{1}{2}n^{-1}\mathbf{u}^\intercal \mathbb{A}\mathbf{u}}{\|n^{-\frac{1}{2}}\mathbf{u}\|} = o_{\mathbb{P}}(1) \tag{1.7}$$

uniformly in $\mathbf{u}$. By the multiplier central limit theorem (Kosorok, 2008, Theorem 10.1),

$$\frac{\mathbb{P}_n^*\{\mathcal{L}(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u})G - \widetilde{\mathcal{L}}(\boldsymbol{\theta}_0)G - n^{-\frac{1}{2}}\mathcal{U}(\boldsymbol{\theta}_0; \mathcal{D})^\intercal \mathbf{u}G\} - \frac{1}{2}n^{-1}\mathbf{u}^\intercal \mathbb{A}\mathbf{u}}{\|n^{-\frac{1}{2}}\mathbf{u}\|} = o_{\mathbb{P}^*}(1) \tag{1.8}$$

uniformly in $\mathbf{u}$. It follows that uniformly for $\boldsymbol{\theta} \in \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq n^{-\frac{1}{2}}\mathbf{u}\}$,

$$\widetilde{\mathcal{L}}^*(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}) - \widetilde{\mathcal{L}}^*(\boldsymbol{\theta}_0) = n^{-\frac{1}{2}}\mathbb{P}_n\{\mathcal{U}(\boldsymbol{\theta}_0; \mathcal{D})G\}\mathbf{u} + \frac{1}{2}n^{-1}\mathbf{u}^\intercal \mathbb{A}\mathbf{u} + o_{\mathbb{P}^*}(n^{-1}\|u\|) \tag{1.9}$$

and thus we may approximate $n\{\widehat{\mathcal{L}}^*(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}) - \widehat{\mathcal{L}}^*(\boldsymbol{\theta}_0)\}$ with $\mathbb{G}_n\{\mathcal{U}(\boldsymbol{\theta}_0; \mathcal{D})G\}\mathbf{u} + \frac{1}{2}\mathbf{u}^\intercal \mathbb{A}\mathbf{u} + n\sum_{j=1}^p p'_{\lambda_{n}^* j}(|\widetilde{\beta}_j^*|) \left(\left| \beta_{0j} + n^{-\frac{1}{2}}u_j \right| - |\beta_{0j}| \right) + o_{\mathbb{P}^*}(\|\mathbf{u}\|^2 + \|\mathbf{u}\|)$.

Now we show the "consistency" of variable selection, i.e., $\mathbb{P}^*(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}^c} = 0) \to 1$ as $n \to \infty$. It suffices to to show that for any constant C and given $\tilde{\boldsymbol{\theta}}_{\mathcal{A}}$ such that $\|\tilde{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0\mathcal{A}}\| = O_{\mathbb{P}^*}(n^{-\frac{1}{2}})$

$$\mathbb{P}^* \left[\text{argmin}_{\|\boldsymbol{\theta}_{\mathcal{A}^c}\| \le C n^{-\frac{1}{2}}} \widehat{\mathcal{L}}^* \left\{ \left(\tilde{\boldsymbol{\theta}}_{\mathcal{A}}^\mathsf{T}, \boldsymbol{\theta}_{\mathcal{A}^c}^\mathsf{T} \right)^\mathsf{T} \right\} = 0 \right] \to 1. \tag{1.10}$$

Let $\tilde{\mathbf{u}}_{\mathcal{A}}$ and $\mathbf{u}_{\mathcal{A}^c}$ denote $n^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0\mathcal{A}})$ and $n^{\frac{1}{2}}\boldsymbol{\theta}_{\mathcal{A}^c}$, respectively. It follows from (1.9)

$$\begin{aligned} & n \left[\widehat{\mathcal{L}}^* \left\{ \left(\boldsymbol{\theta}_{0\mathcal{A}}^\mathsf{T} + n^{-\frac{1}{2}} \tilde{\mathbf{u}}_{\mathcal{A}}^\mathsf{T}, n^{-\frac{1}{2}} \mathbf{u}_{\mathcal{A}^c}^\mathsf{T} \right)^\mathsf{T} \right\} - \widehat{\mathcal{L}}^* \left\{ \left(\boldsymbol{\theta}_{0\mathcal{A}}^\mathsf{T} + n^{-\frac{1}{2}} \tilde{\mathbf{u}}_{\mathcal{A}}^\mathsf{T}, 0^\mathsf{T} \right)^\mathsf{T} \right\} \right] \qquad (1.11) \\ = & \left[\mathbb{G}_n \{ \mathcal{U}(\boldsymbol{\theta}_0; \mathcal{D})_{\mathcal{A}^c}^\mathsf{T} G \} + \tilde{\mathbf{u}}_{\mathcal{A}}^\mathsf{T} \mathbb{A}_{12} \right] \mathbf{u}_{\mathcal{A}^c} + \frac{1}{2} \mathbf{u}_{\mathcal{A}^c}^\mathsf{T} \mathbb{A}_{22} \mathbf{u}_{\mathcal{A}^c} + n \sum_{j \in \mathcal{A}^c} p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|) |n^{-\frac{1}{2}} u_j| \\ & + o_{\mathbb{P}^*}(\|\mathbf{u}_{\mathcal{A}^c}\|^2 + \|\mathbf{u}_{\mathcal{A}^c}\|) = \sum_{j \in \mathcal{A}^c} n^{\frac{1}{2}} p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|) \, |u_j| + R_n(\mathbf{u}_{\mathcal{A}^c}). \qquad (1.12) \end{aligned}$$

where $\sup_{\|\mathbf{u}_{\mathcal{A}^c}\| \le C} R_n(\mathbf{u}_{\mathcal{A}^c}) / \left(\|\mathbf{u}_{\mathcal{A}^c}\|^2 + \|\mathbf{u}_{\mathcal{A}^c}\| \right) = o_{\mathbb{P}^*}(1)$. Zou and Li (2008) consider the limiting behavior of $n^{\frac{1}{2}} p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|)$ for SCAD and L_q penalties in their proof of the oracle properties of the one-step LLA estimator. They show that for both cases, when $j \in \mathcal{A}^c$, $n^{\frac{1}{2}} p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|) \to_p \infty$. Additionally, for the ALASSO penalty, $n^{\frac{1}{2}} p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|) = n^{-\frac{1}{2}} \lambda_n |n^{\frac{1}{2}} \widetilde{\beta}^*_j|^{-1}$, when $j \in \mathcal{A}^c$, we have $n^{-\frac{1}{2}} \lambda_n \to \infty$ and $|n^{\frac{1}{2}} \widetilde{\beta}^*_j| = O_{\mathbb{P}^*}(1)$. Hence, for all three types of penalties, $n^{\frac{1}{2}} p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|) \to_p \infty$. Thus, for any $\epsilon > 0$, there exist $C_1 > C_0 > 0$ and N_0 such that $\mathbb{P}^* \left\{ \sum_{j \in \mathcal{A}^c} n^{\frac{1}{2}} p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|) \, |u_j| \ge C_1 \sum_{j \in \mathcal{A}^c} |u_j| \right\} \ge 1 - \epsilon$ and $\mathbb{P}^* \left\{ C_0 \sum_{j \in \mathcal{A}^c} |u_j| \ge |R_n(\mathbf{u}_{\mathcal{A}^c})| \right\} \ge 1 - \epsilon$ for $\|\mathbf{u}_{\mathcal{A}^c}\| \le C$ and $n \ge N_0$. This implies that with probability greater than $1 - 2\epsilon$, $n \left[\widehat{\mathcal{L}}^* \left\{ \left(\tilde{\boldsymbol{\theta}}_{\mathcal{A}}^\mathsf{T}, n^{-\frac{1}{2}} \mathbf{u}_{\mathcal{A}^c}^\mathsf{T} \right)^\mathsf{T} \right\} - \widehat{\mathcal{L}}^* \left\{ \left(\tilde{\boldsymbol{\theta}}_{\mathcal{A}}^\mathsf{T}, 0^\mathsf{T} \right)^\mathsf{T} \right\} \right] \ge (C_1 - C_0) \sum_{j \in \mathcal{A}^c} |u_j| \ge 0$, which implies (1.10).

Lastly, we will justify the oracle property of $\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}}$. Since $\mathbb{P}^*(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}^c} = 0) \to 1$, $\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}}$ can be considered as the minimizer of $\widehat{\mathcal{L}}^*_{\mathcal{A}}(\boldsymbol{\theta}_{\mathcal{A}}) = \widehat{\mathcal{L}}^* \{ (\boldsymbol{\theta}_{\mathcal{A}}^\mathsf{T}, 0^\mathsf{T})^\mathsf{T} \}$. Following the approach

of Zou (2006), we consider the reparametrization

$$\begin{aligned}&\widehat{\mathcal{L}}^*_{\mathcal{A}}(\boldsymbol{\theta}_{0\mathcal{A}} + n^{-\frac{1}{2}}\mathbf{u}_{\mathcal{A}})\\ &= \mathbb{P}_n\mathcal{L}\left\{\left(\boldsymbol{\theta}^{\intercal}_{0\mathcal{A}} + n^{-\frac{1}{2}}\mathbf{u}^{\intercal}_{\mathcal{A}}, 0^{\intercal}\right)^{\intercal}, \mathcal{D}_i\right\} G_i + \sum_{j\in\mathcal{A}} p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|)\left|\beta_{0j} + n^{-\frac{1}{2}}u_j\right| \quad (1.13)\end{aligned}$$

Let $\hat{\mathbf{u}}^{(n)}_{\mathcal{A}} = \arg\min_{\mathbf{u}_{\mathcal{A}}} \widehat{\mathcal{L}}^*_{\mathcal{A}}(\boldsymbol{\theta}_{0\mathcal{A}} + n^{-\frac{1}{2}}\mathbf{u}_{\mathcal{A}})$. Note $\hat{\mathbf{u}}^{(n)}_{\mathcal{A}} = n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}} - \boldsymbol{\theta}_{0\mathcal{A}})$ is also the minimizer of $V^*_n(\mathbf{u}_{\mathcal{A}}) \equiv \widehat{\mathcal{L}}^*_{\mathcal{A}}(\boldsymbol{\theta}_{0\mathcal{A}} + n^{-\frac{1}{2}}\mathbf{u}_{\mathcal{A}}) - \widehat{\mathcal{L}}^*(\boldsymbol{\theta}_0)$, as $\widehat{\mathcal{L}}^*(\boldsymbol{\theta}_0)$ is a constant. Again, it follows from (1.9)

$$\begin{aligned}V^*_n(\mathbf{u}_{\mathcal{A}}) &= n^{\frac{1}{2}}\mathbf{u}^{\intercal}_{\mathcal{A}}\mathbb{P}_n\{\mathcal{U}_{\mathcal{A}}(\boldsymbol{\theta}_0, \mathcal{D})G\} + \frac{1}{2}\mathbf{u}^{\intercal}_{\mathcal{A}}\mathbb{A}_{11}\mathbf{u}_{\mathcal{A}}\\ &\quad + n\sum_{j\in\mathcal{A}} p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|)\left(\left|\beta_{0j} + n^{-\frac{1}{2}}u_j\right| - |\beta_{0j}|\right)\\ &\quad + o_{\mathbb{P}^*}(\|\mathbf{u}_{\mathcal{A}}\|^2 + \|\mathbf{u}_{\mathcal{A}}\|) \quad (1.14)\end{aligned}$$

To examine the limiting behavior of the third term of $V^*_n(\mathbf{u})$, we have $\beta_{0j} \neq 0$, $n^{\frac{1}{2}}(|\beta_{j0} + n^{-\frac{1}{2}}u_j| - |\beta_{j0}|) \to_p u_j \operatorname{sgn}(\beta_{0j})$, since $j \in \mathcal{A}$. Also, as Zou and Li (2008) proved in their appendix, $n^{\frac{1}{2}}p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|) \to_p 0$ when $j \in \mathcal{A}$ for the SCAD and L_q penalties. For the ALASSO penalty, $n^{\frac{1}{2}}p'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|) = \lambda_n|\widetilde{\beta}_j|^{-1}$, $\lambda_n \to 0$, and $|\widetilde{\beta}_j|^{-1} \to_p |\beta_{0j}|^{-1}$ for $\beta_{0j} \neq 0$. Therefore, by Slutsky's theorem, we have $np'_{\lambda^*_n j}(|\widetilde{\beta}^*_j|)\left(\left|\beta_{0j} + n^{-\frac{1}{2}}u_j\right| - |\beta_{0j}|\right) = o_{\mathbb{P}^*}(1)$ and

$$V^*_n(\mathbf{u}_{\mathcal{A}}) = \mathbf{u}^{\intercal}_{\mathcal{A}}\mathbb{G}_n\{\mathcal{U}_{\mathcal{A}}(\boldsymbol{\theta}_0, \mathcal{D})G\} + \frac{1}{2}\mathbf{u}^{\intercal}_{\mathcal{A}}\mathbb{A}_{11}\mathbf{u}_{\mathcal{A}} + o_{\mathbb{P}^*}(1 + \|\mathbf{u}_{\mathcal{A}}\|^2 + \|\mathbf{u}_{\mathcal{A}}\|). \quad (1.15)$$

Thus, $\hat{\mathbf{u}}^{(n)}_{\mathcal{A}} = -\mathbb{A}^{-1}_{11}\mathbb{G}_n\{\mathcal{U}_{\mathcal{A}}(\boldsymbol{\theta}_0, \mathcal{D})G\} + o_{\mathbb{P}^*}(1)$. Since $\mathbb{G}_n\{\mathcal{U}_{\mathcal{A}}(\boldsymbol{\theta}_0, \mathcal{D})G\}$ converges to $N(\mathbf{0}, \mathbb{B}_{11})$ in distribution, $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}} - \boldsymbol{\theta}_{0\mathcal{A}}) \to_d N(\mathbf{0}, \mathbb{A}^{-1}_{11}\mathbb{B}_{11}\mathbb{A}^{-1}_{11})$ and $\mathbb{P}^*(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}^C} = 0) \to 1$. Then the perturbed regularized estimator $\widehat{\boldsymbol{\theta}}^*$ is asymptotically normal in the true nonzero parameter set.

Similar arguments as given above, along with the conditional multiplier central limit theorem (Kosorok, 2008, Theorem 10.4), can be used to justify that

the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}}) \mid \mathcal{X}$ approximates that of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Specifically, we can similarly obtain $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0\mathcal{A}}) = -\mathbb{A}_{11}^{-1}\mathbb{G}_n\{\mathcal{U}_{\mathcal{A}}(\boldsymbol{\theta}_0, \mathcal{D})\} + o_{\mathbb{P}}(1)$ and $\mathbb{P}^*(\widehat{\boldsymbol{\theta}}_{\mathcal{A}^C} = 0) \to 1$. Therefore, $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}}) = -\mathbb{A}_{11}^{-1}\mathbb{G}_n\{\mathcal{U}_{\mathcal{A}}(\boldsymbol{\theta}_0, \mathcal{D})(G-1)\} + o_{\mathbb{P}^*}(1)$. Since $-\mathbb{A}_{11}^{-1}\mathbb{G}_n\{\mathcal{U}_{\mathcal{A}}(\boldsymbol{\theta}_0, \mathcal{D})(G-1)\} \mid \mathcal{X} \to_d N(\mathbf{0}, \mathbb{A}_{11}^{-1}\widehat{\mathbb{B}}_{11}\mathbb{A}_{11}^{-1})$ and $\widehat{\mathbb{B}}_{11} \to_{\mathbb{P}} \mathbb{B}_{11}$, $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \mid \mathcal{X}$ and $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0\mathcal{A}})$ converge in distribution to the same limit. Furthermore, $\mathbb{P}^*(\widehat{\boldsymbol{\theta}}^*_{\mathcal{A}^C} = 0 | \mathcal{X}) \to 1$.

1.7.2 Choice of thresholding values $\widehat{p}_{high}$ and $\widehat{p}_{low}$ for confidence regions

We choose values for $\widehat{p}_{high}$ and $\widehat{p}_{low}$ to converge at rates relative to the order of the tuning parameter λ and bounded by the probability that our perturbation samples are set to zero. For illustration, consider the univariate β case with one predictor under orthonormal design. Consider standardized parameters $\widehat{\gamma} = \widehat{\beta}/\sigma$, $\gamma = \beta/\sigma$ and $\widetilde{\lambda}_n = \lambda_n/\sigma^2$, where $\lambda_n \to 0$ and $n^{\frac{1}{2}}\lambda_n \to \infty$. Then

$$\widehat{\gamma} \dot{\sim} N(\gamma_0, n^{-1}), \qquad \widehat{\gamma}_1 = \widehat{\gamma}\left(1 - \frac{\widetilde{\lambda}_n}{|\widehat{\gamma}|^2}\right)_+, \qquad \gamma_1^* \dot{\sim} \gamma^*\left(1 - \frac{\widetilde{\lambda}_n}{|\gamma^*|^2}\right)_+ \tag{1.16}$$

where $\gamma^* \dot{\sim} N(\widehat{\gamma}, 1/n)$. Thus $\gamma_1^* = 0$ with probability

$$\hat{P}_0 = P\{|\gamma^*| < \widetilde{\lambda}_n^{1/2}\} = \Phi\{n^{1/2}(\widetilde{\lambda}_n^{1/2} - \widehat{\gamma})\} - \Phi\{-n^{1/2}(\widetilde{\lambda}_n^{1/2} + \widehat{\gamma})\} \tag{1.17}$$

First consider non-zero parameters. Without loss of generality, assume that $\gamma_0 > 0$. Let $\epsilon = 2\widetilde{\lambda}_n^{\frac{1}{2}}$ and assume that $\gamma_0 > 2\epsilon$. Then

$$\begin{aligned} E(\widehat{P}_0) =& E\left[\Phi\left\{n^{\frac{1}{2}}(\widetilde{\lambda}_n^{\frac{1}{2}} - \widehat{\gamma})\right\} - \Phi\left\{n^{\frac{1}{2}}(-\widetilde{\lambda}_n^{\frac{1}{2}} - \widehat{\gamma})\right\}\right] && (1.18)\\ \leq& \left[\Phi\left\{n^{\frac{1}{2}}(\widetilde{\lambda}_n^{\frac{1}{2}} - \epsilon)\right\} - \Phi\left\{n^{\frac{1}{2}}(-\widetilde{\lambda}_n^{\frac{1}{2}} - \epsilon)\right\}\right] P(\widehat{\gamma} > \epsilon) + P(\widehat{\gamma} \leq \epsilon) && (1.19)\\ \leq& 2\Phi(-n^{\frac{1}{2}}\epsilon) \lessapprox \sqrt{2/\pi}\exp\{-n\widetilde{\lambda}_n\} = \sqrt{2/\pi}\exp\{-n\lambda_n/\sigma^2\} && (1.20) \end{aligned}$$

Thus, we propose to choose the lower bound $\widehat{p}_{low} = \min(0.49, \sqrt{2/\pi}\exp\{-n\lambda_n/(4\sigma^2)\})$ such that $\widehat{p}_{low} \gg \sqrt{2/\pi}\exp(-n\lambda_n/\sigma^2)$. On the other hand, if $\gamma_0 = 0$, then

$$E(1-\widehat{P}_0) = E\left[\Phi\left\{n^{\frac{1}{2}}(-\widetilde{\lambda}_n^{\frac{1}{2}} + \widehat{\gamma})\right\} + \Phi\left\{n^{\frac{1}{2}}(-\widetilde{\lambda}_n^{\frac{1}{2}} - \widehat{\gamma})\right\}\right] \tag{1.21}$$

$$\approx 2\Phi\left\{-n^{\frac{1}{2}}\widetilde{\lambda}_n^{\frac{1}{2}}/\sqrt{2}\right\} \lessapprox \sqrt{2/\pi}\exp\{-n\widetilde{\lambda}_n/4\} = \sqrt{2/\pi}\exp\{-n\lambda_n/(4\sigma^2)\} \tag{1.22}$$

Thus, we chose $\widehat{p}^*_{high} = 1 - \sqrt{2/\pi}\exp(-n\lambda_n/\sigma^2)$ such that $\widehat{p}^*_{high} \gg 1 - \sqrt{2/\pi}\exp\{-n\lambda_n/(4\sigma^2)\}$. Note that we chose $\widehat{p}_{low}$ and $\widehat{p}_{high}$ such that $\widehat{p}_{low}$ goes to 0 at a much slower rate than $\widehat{P}_0$ for $\gamma_0 \neq 0$. On the other hand, when $\gamma_0 = 0$, $\widehat{p}^*_{high}$ goes to 1 at a much faster rate than $\widehat{P}_0$ and thus $\widehat{P}_0 > \widehat{p}_{high} = \min(1-\alpha, \widehat{p}^*_{high})$ occurs with probability approaching 1 as $n \to \infty$, for any fixed $\alpha > 0$. Consequently, this indicates a strong evidence of $\gamma = 0$ when $\widehat{P}_0 > \widehat{p}_{high}$. When σ is unknown, it is replaced with a consistent estimate $\widehat{\sigma}$.

1.7.3 Justification of highest density region and bias estimate

For $j \in \mathcal{A}^C$, $\mathbb{P}^*(\widehat{\beta}^*_j = 0) \to 1$ and thus for any $\alpha > 0$, $\mathbb{P}^*(\widehat{\mathcal{P}}_{0j} > \alpha) \to 1$, and $\mathbb{P}(\widehat{\mathcal{P}}_{0j} < \widehat{p}_{high}) + \mathbb{P}(\widehat{\mathcal{P}}_{0j} < \widehat{p}_{low}) \to 0$. Hence, $\mathbb{P}^*(0 \in \mathrm{CR}^{*\mathrm{HDR}}_j) \to 1$ and so we include $\{0\}$ in our CR when $\widehat{\mathcal{P}}_{0j} > \widehat{p}_{low}$ and the coverage of $\mathrm{CR}^{*\mathrm{HDR}}_j$ converges to 1 when $\boldsymbol{\beta}_{0j} = 1$. For $j \in \mathcal{A}$, $\widehat{\mathcal{P}}_{0j} \to_p 0$, and our estimates converge to a continuous distribution, specifically $n^{\frac{1}{2}}(\widehat{\beta}^*_j - \widehat{\beta}_j) \mid \mathcal{X} \to_d N(0, \sigma_j^2)$, where σ_j^2 is the asymptotic variance of $n^{\frac{1}{2}}(\widehat{\beta}_j - \beta_{0j})$. It follows that $\sup_x |n^{-\frac{1}{2}} f^*_j(\widehat{\beta}_{0j} + n^{-\frac{1}{2}}x) - \phi_{\sigma_j}(x)| \to_p 0$ where $\phi_\sigma(x) = \phi(x/\sigma)/\sigma$ and $\phi(\cdot)$ is the density function of the standard normal. Therefore, $\sup_\beta |n^{-\frac{1}{2}} f^*_j(\beta) - \phi_{\sigma_j}\{n^{\frac{1}{2}}(\beta - \widehat{\beta}_{0j})\}| \to_p 0$ and $n^{-\frac{1}{2}}\widehat{c}_3 \to_p c_{30}$, where c_{30}

is the solution to $\int I\{\phi_{\sigma_j}(x) > c_{30}\}\phi_{\sigma_j}(x)dx = 1-\alpha$. It follows that the coverage of our CR converges to nominal levels since, with respect to probability measure $\mathbb{P}^*$, $\text{pr}(\beta_{0j} \in \text{CR}_j^{*\text{HDR}}) = \text{pr}\left\{f_j^*(\beta_{0j}) \geq \widehat{c}_3\right\} + o_{\mathbb{P}^*}(1) = \text{pr}\left\{n^{-\frac{1}{2}} f_j^*(\beta_{0j}) \geq n^{-\frac{1}{2}}\widehat{c}_3\right\} + o_{\mathbb{P}^*}(1) = \text{pr}\left[\phi_{\sigma_j}\{n^{\frac{1}{2}}(\beta_{0j} - \widehat{\beta}_{0j})\} \geq c_{30}\right] + o_{\mathbb{P}^*}(1) \to 1-\alpha$.

Here we define our bias corrected estimator for β_{0j},

$$\widehat{\beta}_j^{BC} = \widehat{\beta}_j + I(\widehat{\beta}_j \neq 0)\widehat{\text{bias}}_j, \tag{1.23}$$

where $\widehat{\text{bias}}_j = \left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\beta}_{j,m}^*\right)(-1)^{I\left[\sum_{m=1}^{M}\{I(\widehat{\beta}_{j,m}^*>0)-I(\widehat{\beta}_{j,m}^*<0)\}<0\right]}\left(\widehat{\mathbb{A}}_\lambda^{-1}\right)_{jj}/\{n\max(|\widehat{\xi}_{7.5}|, |\widehat{\xi}_{97.5}|)\}$, $\widehat{\mathbb{A}}_\lambda = n^{-1}\left(\mathbf{X}_{\widehat{\mathcal{A}}}^{\intercal}\mathbf{X}_{\widehat{\mathcal{A}}} + n^{-\frac{1}{2}}\lambda_n \text{diag}\left\{1/\widetilde{\beta}_j^2\right\}_{j=1}^{p}\right)$ and $\widehat{\xi}_r$ is the r percentile of $\{\widetilde{\beta}_{j,m}^*, m = 1,\ldots M\}$. We estimate $\mathbb{A}$ for ALASSO with $\widehat{\mathbb{A}}_\lambda$ following the methods of Cai et al. (2009) where a stabilized estimate of the covariance of coefficients from an accelerated failure time model is used.

1.8 Appendix B: Model selection

1.8.1 Selection of λ with Bayes Information Criterion

In Section 2, we suggest choosing the tuning parameter λ_n by minimizing the BIC. Here we explicitly present the BIC for the linear regression objective function and ALASSO penalty that we utilized in the simulations and data example in Sections 3 and 4. First, assume that the data has been centered so there is no intercept. We implement a least squares approximation of the likelihood for BIC(λ) as in Wang and Leng (2007). For a given λ,

$$\text{BIC}(\lambda) = (\widehat{\boldsymbol{\beta}}(\lambda) - \widetilde{\boldsymbol{\beta}})^T\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}(\widehat{\boldsymbol{\beta}}(\lambda) - \widetilde{\boldsymbol{\beta}}) + \widehat{q}_\lambda\omega_n \tag{1.24}$$

where $\widehat{\boldsymbol{\beta}}(\lambda)$ minimizes the least squares objective function $\widehat{\mathcal{L}}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \sum_{j=1}^{p} p'_\lambda(|\widetilde{\beta}_j|)|\beta_j|$, based on (1.1), $\widehat{\boldsymbol{\Sigma}}_\lambda^{-1} = (\widehat{\sigma}^2 n)^{-1}\{\mathbf{X}^T\mathbf{X} + \lambda \text{ diag}\{I(\widehat{\beta}_j(\lambda) \neq 0)/|\widetilde{\beta}_j\widehat{\beta}_j(\lambda)|\}_{j=1}^{p}\}$ is a stabilized estimate of $\boldsymbol{\Sigma}$ similar to that in Zou (2006), $\widehat{\sigma}^2$ is the consistent estimate of σ from the linear regression model based on the residual variance, and $\widehat{q}_\lambda$ estimates the degrees of freedom of ALASSO with the number of nonzero elements of $\widehat{\boldsymbol{\beta}}(\lambda)$ (Zou et al., 2007). We choose $\omega_n = \min(n^{0.1}, \log(n))$ because numerical results suggest that $\log(n)$ is much greater than $n^{0.1}$ and leads to excessive shrinkage of moderately sized parameters in finite sample.

Risk classification with an adaptive naive nayes kernel machine model

Jessica Minnier[1], Ming Yuan[2], Jun Liu[3], and Tianxi Cai[1]

[1]Department of Biostatistics
Harvard School of Public Health

[2]School of Industrial and Systems Engineering
Georgia Institute of Technology

[3]Department of Statistics
Harvard University

2.1 Introduction

Accurate and individualized prediction of risk plays a central role in successful disease prevention and treatment selection. Recent advancement in biological and genomic research has led to the discovery of a vast number of new markers associated with disease outcomes. For example, gene expression analyses have identified molecular subtypes that are associated with differential prognosis and response to treatment for breast cancer patients (Perou et al., 2000; Dent et al., 2007). For non-small cell lung cancer patients, several biological markers including cyclin E and Ki-67 were shown to be highly predictive of patient survival (Dosaka-Akita et al., 2001). These new discoveries hold great potential for improving the prediction of clinical outcomes, and may lead to personalized, tailored medicine. To realize the goals of personalized medicine, significant efforts have been made towards building risk prediction models. For example, statistical models for predicting individual risk have been developed for various types of diseases (Gail et al., 1989; Chen et al., 2006; Thompson et al., 2006; Cassidy et al., 2008; Wolf et al., 1991; D'Agostino et al., 1994). However, these models, largely based on traditional clinical risk factors, have limitations in their clinical utilities (Spiegelman et al., 1994; Gail and Costantino, 2001; Vasan, 2006). For example, the predictive accuracy as measured by the C-statistics (Pepe, 2003) was only about 0.70 for the Framingham stroke models (Wolf et al., 1991; D'Agostino et al., 1994) and about 0.60 for the breast cancer Gail model (Gail et al., 1989).

To improve risk prediction for complex diseases, incorporating genotype information into disease risk prediction has been considered an eventuality of modern molecular medicine (Yang et al., 2003; Wray et al., 2008; Johansen and Hegele, 2009; Janssens and van Duijn, 2008). Microarray, genome-wide association studies

(GWAS) as well as next generation sequencing studies provide attractive mechanisms for identifying important genetic markers for complex diseases (McCarthy et al., 2008; Pearson and Manolio, 2008; Mardis, 2008). Despite the initial success of GWAS, these studies focus primarily on the discovery of genetic variants associated with risk. A common approach to incorporate genotype information into risk prediction is to perform genome-wide univariate analysis to identify genetic markers associated with disease risk and then construct a genetic score from the total number of risk alleles or sum of log expression levels. Such a genetic score is then included as a new variable in the risk prediction model and assessed for its incremental value in risk prediction. However, adding such simple risk scores to the prediction model has led to little improvement in risk prediction accuracy (Gail, 2008; Meigs et al., 2008; Purcell et al., 2009; Lee et al., 2012). This is in part due to the fact that non-linear and interactive effects that may contribute to disease risk have not yet been identified or incorporated. (Marchini et al., 2005; McKinney et al., 2006; Wei et al., 2009). Furthermore, existing findings have shown that common genetic variants often explain a small portion of genetic heritability of complex diseases and suggest that numerous genes may simultaneously affect the disease risk (Visscher et al., 2008; Paynter et al., 2010; Wacholder et al., 2010; Machiela et al., 2011; Makowsky et al., 2011). Therefore, to achieve optimal accuracy, one must incorporate such complex effects from multiple genes into the new risk prediction model.

Statistical procedures for combining markers to improve risk prediction have been proposed for linear additive effects with a small number of markers (Su and Liu, 1993; McIntosh and Pepe, 2002; Pepe et al., 2006). However, relatively little statistical research has been done on risk prediction in the presence of high dimensional markers with complex non-linear interactive effects. Current literature on studying interactive

effects focuses primarily on testing for the significance of interactions (Umbach and Weinberg, 1997; Yang and Khoury, 1997; Chatterjee and Carroll, 2005; Murcray et al., 2009). Traditional statistical methods that include explicit interaction terms in regression are not well suited for detecting or quantifying such interactive and non-linear effects, especially when the number of predictors is not very small and when higher order and non-linear interactions are present. To overcome such difficulties, we propose to employ a kernel machine (KM) regression framework which has emerged in the last decade as a powerful technique to incorporate complex effects (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002). Recently, statistical procedures for making inference about model parameters under KM regression framework have been proposed (Li and Luan, 2003; Liu et al., 2007, 2008). The KM models implicitly specify the underlying complex functional form of covariate effects via knowledge-based similarity measures that define the distance between two sets of covariates. These procedures, while useful in capturing non-linear effects, may not be efficient when the underlying model is too complex. The lack of efficiency is even more pronounced when the number of candidate markers is large, with the possibility that many such markers are unrelated to the risk.

To achieve a good balance between model complexity and estimation efficiency, we propose a multi-stage adaptive estimation procedure when the genomic markers are partitioned into M gene-sets based on prior knowledge. In the first stage, by imposing a blockwise naive bayes KM (NBKM) model, the marker effects within a gene-set are allowed to be complex and interactive while the total effects from the M gene-sets are assumed to be aggregated additively. Within each gene-set, we propose to improve the estimation via an KM principal component analysis (PCA) (Schölkopf and Smola, 2002; Bengio et al., 2004; Braun, 2005) which effectively reduces the

degree of freedom. In the second stage, we recalibrate our estimates adaptively via a blockwise variable selection procedure to account for the fact that some of the gene-sets may be unrelated to the risk and the model imposed in the first stage may not be optimal. The NBKM model is described in section 2 and the detailed procedures for model estimations are given in section 3. Procedures for assessing the predictive accuracy of the resulting risk score are given in section 4. In section 5, we first provide results from simulation studies illustrating the performance of our proposed procedures and compare them to some of the existing procedures. Then, applying our methods to a GWAS of type I diabetes (T1D) collected by Welcome Trust Case Control Consortium (WTCCC), we obtain a genetic risk score classifying T1D and evaluate its accuracy in classifying the T1D disease status. Some closing remarks are given in section 6.

2.2 Naive-bayes kernel machine (NBKM) model

Let Y denote the binary outcome of interest with $Y = 1$ being diseased and $Y = 0$ being non-diseased. Suppose there are M distinct gene-sets available for predicting Y and we let $\mathbf{Z}^{(m)}$ denote the vector of genetic markers in the mth set. The gene-sets can be created via biological criteria such as genes, pathways, or linkage disequilibrium (LD) blocks. Let $\mathbf{Z}^{(\bullet)} = (\mathbf{Z}^{(1)^\intercal}, ..., \mathbf{Z}^{(M)^\intercal})^\intercal$ denote the entire vector of genetic markers from all M sets. Assume that data for analysis consist of n independent and identically distributed random vectors, $\{(Y_i, \mathbf{Z}_i^{(\bullet)}, l = 1, ..., M), i \in \mathscr{D}\}$, where $\mathscr{D} = \{1, ..., n\}$ indexes all subjects the entire dataset. Throughout, we use the notation $\|\cdot\|_1$ and $\|\cdot\|_2$ to denote the L_1 and L_2 vector norm.

To construct a prediction model for Y based on $\mathbf{Z}^{(\bullet)}$, we impose a working *Naive*

Bayes (NB) assumption that $\{\mathbf{Z}^{(m)}, m = 1, \ldots, M\}$ are independent of each other conditional on Y. Under this assumption, it is straightforward to see that

$$\text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(\bullet)}) = a + \sum_{m=1}^{M} \text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(m)}), \tag{2.1}$$

and thus $\text{pr}(Y = 1 \mid \mathbf{Z}^{(\bullet)})$ can be approximated by first approximating $\text{pr}(Y = 1 \mid \mathbf{Z}^{(m)})$ using data from the mth gene-set only. To estimate $\text{pr}(Y = 1 \mid \mathbf{Z}^{(m)})$, we assume a logistic KM model

$$\text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(m)}) = a^{(m)} + h^{(m)}(\mathbf{Z}_i^{(m)}) \tag{2.2}$$

where $h^{(m)}(\cdot) \in \mathcal{H}_k^{(m)}$ is an unknown centered smooth function and the functional space $\mathcal{H}_k^{(m)}$ is implicitly specified by a positive definite kernel function $k(\cdot, \cdot)$.

For any pair of genetic marker vectors $(\mathbf{z}_1, \mathbf{z}_2)$, $k(\mathbf{z}_1, \mathbf{z}_2)$ measures the similarity or distance between $\mathbf{z}_1$ and $\mathbf{z}_2$. Examples of kernel functions under investigation include

(i) polynomial kernel: $k_{\text{POLY}}(\mathbf{z}_1, \mathbf{z}_2; d) = (1 + \mathbf{z}_1^{\intercal}\mathbf{z}_2)^d$ corresponding to d-way multiplicative interactive effects;

(ii) IBS kernel for genetic markers: $k_{\text{IBS}}(\mathbf{z}_1, \mathbf{z}_2) = \sum_{l=1}^{p} \text{IBS}(z_{1l}, z_{2l})$, where $\text{IBS}(z_{1l}, z_{2l})$ represents the number of alleles shared identity by state;

(iii) Gaussian Kernel: $k_{\text{GAU}}(\mathbf{z}_1, \mathbf{z}_2) = \exp\{-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\rho\}$ which allows for complex non-linear smooth effects, where ρ is a tuning parameter.

These kernel functions have been advocated as effective in capturing interactive effects (Schölkopf and Smola, 2002; Kwee et al., 2008). The choice of k directly impacts the complexity and predictive performance of the model and should be selected based

on the biological knowledge and empirical evidence of the relationship between $\mathbf{Z}$ and Y.

Under the naive bayes kernel machine (NBKM) model assumptions given in (2.1) and (3.5), the conditional likelihood of Y given $\mathbf{Z}^{(\bullet)}$ is a monotone function of $\sum_{m=1}^{M} h^{(m)}(\mathbf{Z}^{(m)})$. Therefore, $\sum_{m=1}^{M} h^{(m)}(\mathbf{Z}^{(m)})$ is the optimal risk score of $\mathbf{Z}^{(\bullet)}$ for classifying Y in the sense that $\sum_{m=1}^{M} h^{(m)}(\mathbf{Z}^{(m)})$ achieves the highest ROC curve among all risk scores determined by $\mathbf{Z}^{(\bullet)}$ (McIntosh and Pepe, 2002). It follows that the optimal risk score can be estimated by separately fitting the mth KM model (3.5) to data from the mth gene-set: $\{(Y_i, \mathbf{Z}_i^{(m)}), i = 1, ..., n\}$.

2.3 Model estimation under the NBKM model

2.3.1 Kernel PCA estimation for modeling the mth gene-Set

To estimate $h^{(m)}$ based on data from the mth gene-set, we note that by Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), any $h^{(m)}(\mathbf{z}) \in \mathcal{H}_k^{(m)}$ has a *primal representation*,

$$h^{(m)}(\mathbf{z}) = \sum_{j=1}^{\infty} \beta_j^{(m)} \psi_j^{(m)}(\mathbf{z}), \tag{2.3}$$

where $\{\beta_j^{(m)}\}$ are the square summable unknown coefficients, $\psi_j^{(m)}(\mathbf{z}) = \sqrt{\lambda_j^{(m)}}\phi_j^{(m)}(\mathbf{z})$, $\{\lambda_j^{(m)}\}$ and $\{\phi_j^{(m)}\}$ are the eigenvalues and eigenfunctions of k under the probability measure $\mathcal{P}_{\mathbf{Z}^{(m)}}$, where $\lambda_1^{(m)} \geq \lambda_2^{(m)} \geq \cdots \geq 0$ and $\mathcal{P}_{\mathbf{Z}^{(m)}}$ is the distribution of $\mathbf{Z}^{(m)}$. For finite samples, a suitable approach to incorporate the potentially large number of parameters associated with $h^{(m)}$ is to maximize a penalized logistic likelihood function with the penalty accounting for the smoothness of $h^{(m)}$. However, since the forms of the basis functions for h, $\{\psi_j^{(m)}(\mathbf{z})\}$, are intractable in general, it is not

feasible to directly use the primal representation to estimate $h^{(m)}$. On the other hand, by the representer theorem (Kimeldorf and Wahba, 1971), the maximum penalized likelihood estimator for $h^{(m)} \in \mathcal{H}_k^{(m)}$ must admit a *dual representation*,

$$h^{(m)}(\mathbf{z}) = \sum_{j=1}^{n} \alpha_j^{(m)} k(\mathbf{z}, \mathbf{Z}_j^{(m)}), \tag{2.4}$$

where $\{\alpha_j^{(m)}\}$ are the unknown regression parameters. An estimator of $(a^{(m)}, \boldsymbol{\alpha}^{(m)})$ can be obtained as the maximizer of the penalized log-likelihood function as given in Liu et al. (2008),

$$\mathcal{L}^{(\mathrm{D})}(a, \boldsymbol{\alpha}; \mathbf{K}_{n(m)}) = \mathbf{Y}^{\intercal} \log g(a+\mathbf{K}_{n(m)}\boldsymbol{\alpha}) + (1-\mathbf{Y})^{\intercal} \log\{1-g(a+\mathbf{K}_{n(m)}\boldsymbol{\alpha})\} - \tau \boldsymbol{\alpha}^{\intercal}\mathbf{K}_{n(m)}\boldsymbol{\alpha}, \tag{2.5}$$

where $\mathbf{Y} = (Y_1, ..., Y_n)^{\intercal}$, $g(\cdot) = \mathrm{logit}^{-1}(\cdot)$, $\mathbf{K}_{n(m)} = n^{-1}[k(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)})]_{1 \le i,j \le n}$, and τ is a tuning parameter controlling the amount of regularization.

The above Liu et al. (2008) estimator may not be efficient due to the high dimensionality in the parameter space and could be numerically challenging to obtain when the sample size n and hence the dimension of $\boldsymbol{\alpha}$ is not small, as in many GWAS settings. To improve the estimation precision, we propose the use of the kernel PCA (Schölkopf and Smola, 2002; Bengio et al., 2004; Braun, 2005) where only the PCs with large eigenvalues are included for estimation. When the eigenvalues $\{\lambda_j^{(m)}\}$ decay quickly, the feature space $\mathcal{H}_k^{(m)}$ may be approximated well with the space spanned by the leading eigenfunctions and $\{\beta_j^{(m)}\sqrt{\lambda_j^{(m)}}\}$ may also decay quickly. Due to the bias and variance trade-off, the estimation of $h^{(m)}$ may be improved by employing the approximated feature space. To this end, we apply a singular value decomposition to $\mathbf{K}_{n(m)}$, and denote the non-decreasing eigenvalues by $(l_1^{(m)}, \ldots.., l_n^{(m)})$ and the corresponding eigenvectors by $(\mathbf{u}_1^{(m)}, ..., \mathbf{u}_{r_n}^{(m)}, ..., \mathbf{u}_n^{(m)})$. Let $r_n^{(m)}$ be the smallest r such that $\sum_{i=1}^{r} l_i^{(m)} / \sum_{i=1}^{n} l_i^{(m)} \ge \wp$, where $\wp \in (0,1)$ is a pre-specified proportion. The

kernel PCA approximation to $\mathbf{K}_{n(m)}$ corresponding to these $r_{n(m)}$ eigenvalues is then $\mathbf{K}_{n(m)}^{[r_{n(m)}]} = \mathbb{U}_{(m)}\mathbb{D}_{(m)}\mathbb{U}_{(m)}^{\mathsf{T}}$, where

$$\mathbb{U}_{(m)} = \left[\mathbf{u}_1^{(m)}, ..., \mathbf{u}_{r_{n(m)}}^{(m)}\right], \quad \text{and} \quad \mathbb{D}_{(m)} = \text{diag}\left\{l_1^{(m)}, ..., l_{r_{n(m)}}^{(m)}\right\} \tag{2.6}$$

With the kernel PCA approximation, one may estimate $(a^{(m)}, \boldsymbol{\alpha}^{(m)})$ as the maximizer of $\mathcal{L}^{(\text{D})}(a, \boldsymbol{\alpha}; \mathbf{K}_{n(m)}^{[r_{n(m)}]})$, where $\mathcal{L}^{(\text{D})}(a, \boldsymbol{\alpha}; \mathbf{K}_{n(m)}^{[r_{n(m)}]})$ is obtained by replacing $\mathbf{K}_{n(m)}$ in $\mathcal{L}^{(\text{D})}(a, \boldsymbol{\alpha}; \mathbf{K}_{n(m)})$ with $\mathbf{K}_{n(m)}^{[r_{n(m)}]}$. However, since $\mathbf{K}_{n(m)}^{[r_{n(m)}]}$ is singular, such a maximization does not have a unique solution and thus is unstable. We propose a reparametrization with

$$\boldsymbol{\beta}^{(m)} = \widetilde{\boldsymbol{\Psi}}_{(m)}^{\mathsf{T}}\boldsymbol{\alpha}^{(m)}, \quad \text{where} \quad \widetilde{\boldsymbol{\Psi}}_{(m)} = \mathbb{U}_{(m)}\text{diag}\left\{\sqrt{l_1^{(m)}}, ..., \sqrt{l_{r_n}^{(m)}}\right\}. \tag{2.7}$$

This reparameterization essentially links the dual representation back to the primal representation since the observed eigenvalues and eigenvectors of $\mathbf{K}_{n(m)}$ are approximating the corresponding eigenvalues and eigenfunctions of k under the probability measure $\mathcal{P}_{\mathbf{Z}^{(m)}}$. Thus, the kernel PCA approximation along with the reparameterization results in approximating $\{h^{(m)}(\mathbf{Z}_1^{(m)}), ..., h^{(m)}(\mathbf{Z}_n^{(m)})\}^{\mathsf{T}}$ with $\widetilde{\boldsymbol{\Psi}}_{(m)}\boldsymbol{\beta}^{(m)}$. This approach also has computational advantages due to the reduction in the number of unknown parameters from n in the dual form to $r_{n(m)}$ which is often much smaller than n. With the reparametrization, for the training samples, we essentially transform the original covariate matrix $(\mathbf{Z}_1^{(m)}, ..., \mathbf{Z}_n^{(m)})^{\mathsf{T}}$ to $\widetilde{\boldsymbol{\Psi}}_{(m)}$ and estimate $\{h^{(m)}(\mathbf{Z}_1^{(m)}), ..., h^{(m)}(\mathbf{Z}_n^{(m)})\}^{\mathsf{T}}$ as $\widetilde{\boldsymbol{\Psi}}_{(m)}\widehat{\boldsymbol{\beta}}^{(m)}$, where $\{\widehat{a}^{(m)}, \widehat{\boldsymbol{\beta}}^{(m)}\} = \text{argmax}_{a,\boldsymbol{\beta}}\{\mathcal{L}^{(\text{P})}(a, \boldsymbol{\beta}; \widetilde{\boldsymbol{\Psi}}_{(m)})\}$, where

$$\mathcal{L}^{(\text{P})}(a, \boldsymbol{\beta}; \widetilde{\boldsymbol{\Psi}}_{(m)}) = \mathbf{Y}^{\mathsf{T}}\log g(a + \widetilde{\boldsymbol{\Psi}}_{(m)}\boldsymbol{\beta}) + (1 - \mathbf{Y})^{\mathsf{T}}\log\{1 - g(a + \widetilde{\boldsymbol{\Psi}}_{(m)}\boldsymbol{\beta})\} - \tau\|\boldsymbol{\beta}\|_2^2, \tag{2.8}$$

and $\tau \geq 0$ is a tuning parameter that can be selected via criteria such as the AIC or cross-validation, such that $n^{-\frac{1}{2}}\tau \to 0$.

To estimate $\mathbf{H}(\mathbf{z}^{(\bullet)}) = \{h^{(1)}(\mathbf{z}^{(1)}), ..., h^{(M)}(\mathbf{z}^{(M)})\}^{\intercal}$ for a future subject with marker value $\mathbf{Z}^{(\bullet)} = \mathbf{z}^{(\bullet)}$, one may find the transformed covariate in the non-linear feature space via the Nyström method (Rasmussen, 2004) as

$$\widehat{\mathbf{\Psi}}_{(m)}(\mathbf{z}^{(m)}) = n^{-1}\text{diag}\left(\frac{1}{\sqrt{l_1^{(m)}}}, ..., \frac{1}{\sqrt{l_{r_n}^{(m)}}}\right)\mathbb{U}_{(m)}^{\intercal}\left[k(\mathbf{z}^{(m)}, \mathbf{Z}_1^{(m)}), \cdots, k(\mathbf{z}^{(m)}, \mathbf{Z}_n^{(m)})\right]. \tag{2.9}$$

Subsequently, we estimate $h^{(m)}(\mathbf{z}^{(m)})$ as $\widehat{h}^{(m)}(\mathbf{z}^{(m)}) = \widehat{\mathbf{\Psi}}_{(m)}(\mathbf{z}^{(m)})\widehat{\boldsymbol{\beta}}^{(m)}$. In Appendix A, we show that our estimator is root-n consistent for $h(\cdot)$ under the assumption that the reproducible kernel hilbert space $\mathcal{H}_k^{(m)}$ is finite dimensional. This is often a reasonable assumption for GWAS settings since each gene-set has a finite set of single-nucleotide polymorphism (SNP) markers, which can only span a finite dimensional space regardless the choice of kernel.

2.3.2 Combining multiple gene-sets for risk prediction

With the estimated $\widehat{h}^{(m)}$, one may simply classify a future subject with $\mathbf{Z}^{(\bullet)} = \{\mathbf{z}^{(m)}, m = 1, ..., M\}$ based on $\sum_{m=1}^{M}\widehat{h}^{(m)}(\mathbf{z}^{(m)})$ under the naive bayes assumption. However, since some of the gene-sets may not be associated with disease risk, including $\widehat{h}^{(m)}$ from these gene-sets in the model may lead to a decrease in the precision of prediction and risk score estimation. To further improve the precision, we propose to employ an LASSO regularization procedure (Tibshirani, 1996) in the second step to estimate the optimal weight for each individual gene-set. The regularized estimation would assign a weight zero for the non-informative regions while simultaneously providing stable weight estimates for the informative regions. Specifically, based on the synthetic data $\{\mathbf{Y}, \widehat{\mathbb{H}}\}$ constructed from the first step, we re-weight the gene-sets

in the second step by fitting the logistic model

$$\text{logitpr}(Y = 1 \mid \mathbf{Z}^{(\bullet)}) = b_0 + \boldsymbol{\gamma}^\intercal \widehat{\mathbf{H}}(\mathbf{Z}^{(\bullet)}) \tag{2.10}$$

where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_M)^\intercal$, $\widehat{\mathbf{H}}(\mathbf{Z}^{(\bullet)}) = [\widehat{h}^{(1)}(\mathbf{Z}^{(1)}), ..., \widehat{h}^{(M)}(\mathbf{Z}^{(M)})]^\intercal$ and $\widehat{\mathbb{H}} = [\widehat{h}^{(m)}(\mathbf{Z}_i^{(m)})]_{n\times M}$. We obtain a LASSO regularized estimate of $\{b_0, \boldsymbol{\gamma}\}$, as

$$\{\hat{b}, \widehat{\boldsymbol{\gamma}}\} = \underset{b,\boldsymbol{\gamma}}{\text{argmax}} \left\{ \mathcal{L}_{\widehat{\mathbb{H}}}(b, \boldsymbol{\gamma}) - \tau_2 \|\boldsymbol{\gamma}\|_1 \right\}, \tag{2.11}$$

where $\tau_2 \geq 0$ is a tuning parameter such that $n^{-\frac{1}{2}}\tau_2 \to 0$ and $\tau_2 \to \infty$, and

$$\mathcal{L}_{\widehat{\mathbb{H}}}(b, \boldsymbol{\gamma}) = \mathbf{Y}^\intercal \log g(b + \widehat{\mathbb{H}}\boldsymbol{\gamma}) + (1 - \mathbf{Y})^\intercal \log\{1 - g(b + \widehat{\mathbb{H}}\boldsymbol{\gamma})\} \tag{2.12}$$

It is important to note that our estimator $\widehat{\boldsymbol{\gamma}}$ is essentially an adaptive LASSO (Zou, 2006) type estimator since these weights are multiplied with $\widehat{h}^{(m)}(\mathbf{z})$ which are consistent for $h^{(m)}$. As a result, $\widehat{\boldsymbol{\gamma}}$ exhibits the gene-set selection consistency property such that $P(\widehat{\mathcal{A}} = \mathcal{A}) \to 1$ as $n \to \infty$, where $\mathcal{A} = \{m : h^{(m)}(\mathbf{z}) \neq 0\}$ and $\widehat{\mathcal{A}} = \{m : \widehat{\gamma}_m \neq 0\}$. Therefore, this method of estimation consistently includes only informative regions in the prediction model. We show in Appendix B that the proposed adaptively reweighting procedure is consistent in group selection, i.e. $P(\widehat{\mathcal{A}} = \mathcal{A}) \to 1$ in probability as $n \to \infty$.

When such a risk prediction model is formed, it is crucial to assess its ability in discriminating subjects with or without disease. For a given risk score $\mathcal{P}$, the discrimination accuracy can be summarized based on various measures such as the area under the receiver operating characteristic (ROC) curve (AUC) (Swets, 1988; Pepe, 2003). The ROC curve is determined from plotting sensitivity against 1-specificity for all possible cut-offs of the risk score. An AUC of 1 indicates a perfect prediction and 0.5 indicates a random result. Few clinical scores achieve AUCs

ranging > 0.75, and scores with an AUC of 0.95 or greater are considered excellent. Since the number of parameters involved in the training the proposed risk score could be quite large, the AUC should be estimated empirically in an independent validation set. This validation set may be a new data set, or one could set aside a random sample of the data so that $\mathscr{D}$ is partitioned into $\mathscr{D}_t$ and $\mathscr{D}_v$ prior to building the model.

2.3.3 Improved estimation of $\boldsymbol{\gamma}$ via cross-validation

Based on the estimation procedures described in section 2.3, we may estimate the probability of disease for a future subject with $\mathbf{Z}^{(\bullet)}$ under the NBKM as

$$\widetilde{P}(\mathbf{Z}^{(\bullet)}) = g\left\{\hat{b} + \widehat{\boldsymbol{\gamma}}^{\mathsf{T}}\widehat{\mathbb{H}}(\mathbf{Z}^{(\bullet)})\right\}. \tag{2.13}$$

However, training of the KM model for each specific gene-set involves complex models with a potentially large number of effective model parameters, the estimation of $\boldsymbol{\gamma}$ in the second stage may also suffer from instability due to overfitting if we estimate $\boldsymbol{\gamma}$ on the same dataset that we use to estimate $\boldsymbol{\beta}$ for $h(\mathbf{z})$.

To overcome overfitting issues, we propose a K-fold cross-validation procedure to partition the training data $\mathscr{D}_t$ of size n_t into K parts of approximately equal sizes, denoted by $\{\mathscr{D}_{t(k)}, k = 1\ldots, K\}$. For each k we use data *not* in $\mathscr{D}_{t(k)}$ to obtain an estimate for $h^{(m)}$ as $\widehat{h}^{(m)}_{t(-k)}$ based on procedures described in section 2.3.1; and then use those estimates to predict subjects in $\mathscr{D}_{t(k)}$ to obtain $\widehat{\mathbb{H}}_{t(k)} = [\widehat{h}^{(m)}_{t(-k)}(\mathbf{Z}^{(m)}_{t(k)i})]_{\frac{n_t}{K}\times M}$. Then an improved estimate of $\boldsymbol{\gamma}$, denoted by $\widehat{\boldsymbol{\gamma}}_{cv}$, can be obtained via maximization of

$$\sum_{k=1}^{K}\left[\mathbf{Y}^{\mathsf{T}}_{t(k)}\log g(b + \widehat{\mathbb{H}}_{t(k)}\boldsymbol{\gamma}) + (1-\mathbf{Y}_{t(k)})^{\mathsf{T}}\log\{1 - g(b + \widehat{\mathbb{H}}_{t(k)}\boldsymbol{\gamma})\}\right] - \tau_2\|\boldsymbol{\gamma}\|_1, \tag{2.14}$$

This procedure maximizes the use of the training set to estimate $\widehat{\boldsymbol{\gamma}}_{cv}$ while reducing overfitting bias. As shown in the simulation section, this method provides a more accurate estimate of $\boldsymbol{\gamma}$ than using the entire $\mathscr{D}_t$ without cross-validation which leads to overfitting. The consistency of $\widehat{\boldsymbol{\gamma}}_{cv}$ can be established using similar arguments as those given in Appendix E for $\widehat{\boldsymbol{\gamma}}$.

We then use the entire training set $\mathscr{D}_t$ to obtain an estimate of $\mathbf{H}$ as $\widehat{\mathbb{H}}(\mathbf{Z}^{(\bullet)})$ on for an out of sample subject with covariate data $\mathbf{Z}^{(\bullet)}$. The final estimated risk prediction model would thus predict the risk of disease for this new subject as

$$\widehat{P}(\mathbf{Z}^{(\bullet)}) = g\{\widehat{b}_t + \widehat{\boldsymbol{\gamma}}_{cv}^{\intercal}\widehat{\mathbb{H}}(\mathbf{Z}^{(\bullet)})\} \tag{2.15}$$

2.4 Numerical analyses

2.4.1 Type I diabetes GWAS dataset

Type I diabetes (T1D), also known as juvenile-onset diabetes, is a chronic autoimmune disease characterized by insulin deficiency and hyperglycemia due to the destruction of pancreatic islet beta cells. Diagnosis and onset often occurs in childhood. Since the discovery of the association of the disease with the HLA sequence polymorphisms in the late 1980s, the understanding of T1D pathogenesis has advanced with the identification of additional genetic risk factors for the disease (Van Belle et al., 2011). T1D is thought to be triggered by environmental factors in genetically susceptible individuals. However, the proportion of newly diagnosed children with known high-risk genotypes has been decreasing, suggesting that further genetic risk markers have not yet been discovered (Borchers et al., 2010).

Compiling information from a number of large scale genetic studies conducted and published in recent years, the National Human Genome Research Institute (NHGRI) provides an online catalog which lists 75 single nucleotide polymorphisms (SNPs) that have been identified as T1D risk alleles (Hindorff et al., 2009, http://www.genome.gov/gwastudies/ Accessed December 10, 2011) and 91 genes that either contain these SNPs or flank the SNP on either side on the chromosome. Expanding the search to other documented autoimmune diseases (Rheumatoid arthritis, Celiac disease, Crohn's disease, Lupus, Inflammatory bowel disease), the NHGRI lists 375 genes containing or flanking 365 SNPs that have been found to be associated with this class of diseases.

Included among the studies listed in the NHGRI catalog is a large-scale GWAS collected by WTCCC, a group of 50 research groups across the UK that was formed in 2005. The study, detailed in Burton et al. (2007), consists of 2000 T1D cases and 3000 controls of European descent from Great Britain. The control subjects were drawn from the 1958 British Birth Cohort and the UK Blood Services. Approximately 482,509 SNPs were genotyped on an Affymetrix GeneChip 500K Mapping Array Set. We chose to segment the genome on the 22 autosomal chromosomes into gene-sets that include a gene and a flanking region of 20KB on either side of the gene. The WTCCC data we use for analysis includes 350 gene-sets that either contain or lie up- or down-stream of the 365 SNPs that were previously found to be associated with autoimmune diseases. The data includes 40 of the 75 SNPs that were previously found to be associated with T1D. These 350 gene-sets cover 9,256 SNPs present in the WTCCC data.

2.4.2 Simulation studies

We first present results from simulation studies with data generated from the SNP data from the WTCCC study. To assess the performance of our methods, we chose settings that reflect possible types of genetic association with disease risk. For illustrative purposes, we let $\mathbf{Z}^{(\bullet)}$ be the aforementioned $M = 350$ gene-sets. We generated the disease status of 1500 subjects from the logistic regression model, $\text{logit}P(Y = 1|\mathbf{Z}^{(\bullet)}) = \sum_{m=1}^{4} h^{(m)}(\mathbf{Z}^{(m)})$, where we modeled $h^{(m)}(\mathbf{z})$ for $m = 1, \ldots 4$ as linear or nonlinear functions of $\mathbf{Z}^{(m)}$, with varying degrees of complexity. The remaining 346 gene-sets were included as non-informative regions. The labels used in the subsequent tables are denoted in parentheses in the following model descriptions. We present the results from three settings where $h^{(m)}(\mathbf{z})$ for $l = 1, \ldots 4$ are all linear (allL), all nonlinear (allNL), or two linear and two nonlinear functions (LNL). We relegate details about the forms of these functions to Appendix F.

We partition each dataset into a training set of 1000 and a validation set of 500 subjects. We estimate $h^{(m)}(\cdot)$ using the training set by fitting the block specific KM model with either a linear kernel function, k_{LIN}, or an IBS kernel function, k_{IBS}. To compare the performance of our KM PCA approach to the Liu et al. (2008) approach, we obtain estimates by maximizing (2.5) with the full kernel matrix (noPCA) and also based on the PCA approximated likelihood in (2.8) with $\wp = .99$ or .999. When combining information across the M blocks, we use both $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\gamma}}_{cv}$ described in section 2.3 to estimate $\boldsymbol{\gamma}$. We compare our adaptive weighting scheme (ANB) that adaptively estimate $\boldsymbol{\gamma}$ to the purely naive bayes approach where $\boldsymbol{\gamma} = 1$ (NB). Additionally, we compare our methods to models that do not incorporate the block structure of the data by fitting three global modelswith all 9,256 SNPs in the 350

gene-sets: (1) a global KM model with k_{IBS} (gIBS), (2) a global ridge regression model (gRidge), as well as (3) the sure independence screening procedure (SIS) described in Fan and Lv (2008). Lastly, we compare our methods to the weighted sum of the marginal log odds ratios for each of the SNPs (WLGR). The tuning parameter was selected by maximizing the AIC for the ridge regression model in the first stage and via the BIC for the LASSO model in the second stage for combining across blocks. The results are based on 1500 Monte Carlo simulations.

First, we present results on selecting informative blocks via our second stage adaptive estimation of $\boldsymbol{\gamma}$. As shown in Figure 2.1, the informative blocks have nonzero estimated coefficients with high probability for strong signals, though the power to select blocks can be lower for blocks with weaker signals. Noninformative blocks are excluded from the model with very high probability, illustrating the oracle property of $\widehat{\boldsymbol{\gamma}}$ proved in the appendix. In Table 2.1 we see that the method with k_{LIN} selects a larger model on average than the method with k_{IBS} but has a lower probability of selecting the informative gene-sets with nonlinear effects. The method without PCA gives a larger model on average and performs similarly to our method in choosing the correct informative gene-sets. Overall, the best performance in estimation and gene-set selection is seen for models with k_{IBS}.

Table 2.1: Average model size (average number of true blocks selected) from simulation studies for the adaptively weighted gene-set regression model. The true model includes four informative blocks.

$\mathcal{K}$	$\wp$	allL	LNL	NL
IBS	.999	4.1 (2.7)	4.0 (3.0)	4.5 (3.1)
LIN	.999	5.4 (2.8)	5.7 (2.8)	4.4 (2.5)
IBS	noPCA	4.2 (2.7)	4.1 (3.0)	4.5 (3.1)
LIN	noPCA	6.6 (2.9)	7.0 (2.8)	5.3 (2.5)

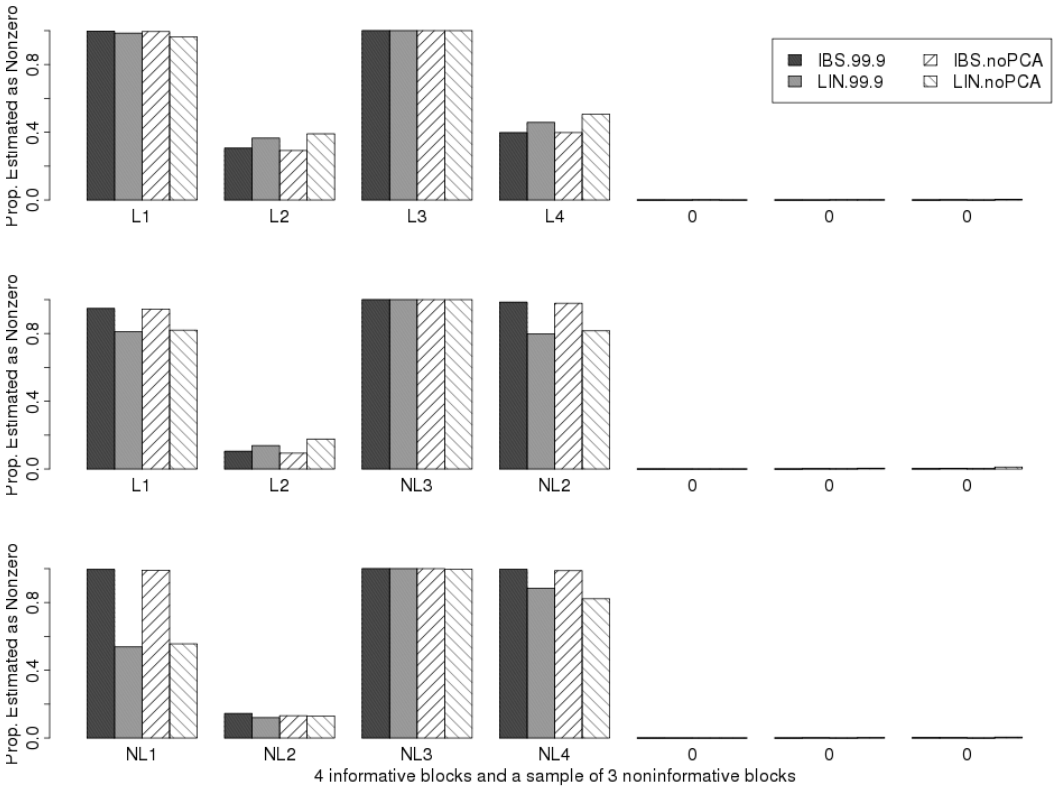

Figure 2.1: Proportion of $\widehat{\gamma}_m$ estimated as nonzero from simulation studies for the adaptively weighted gene-set regression model. Presented are proportions for the four informative blocks as well as a sample of 3 out of the 346 non-informative blocks for each of the three settings representing different types of effects within each informative block (all linear, 2 linear and 2 nonlinear, and all nonlinear).

To compare the methods with respect to predictive performance, we project the model estimates into a validation set of 500 subjects and report the AUC estimates and their standard errors from all models in Table 3.1. The global methods (gRidge, gIBS, SIS, WLGR) generally have substantially worse predictive performances compared to our proposed ANB procedures, suggesting the benefit of taking advantage of blocking combined with adaptive weighting. The benefit of blocking can be also highlighted when we compare results between the ANB procedures and the SIS pro-

cedures. The SIS procedures outperform global ridge and WLGR procedures with higher AUC values, but also have larger standard errors than any other method. Even when all effects are linear and SIS performs fairly well with higher AUC than other global methods as well as NB blockwise methods, we still see substantial improvement in prediction when applying an ANB method with either the linear or IBS kernel. Although both procedures allow for marker selection, the ANB procedures can more effectively estimate the effect of informative blocks and remove the non-informative blocks. When comparing ANB and NB procedures, we see that similar to the global methods, the purely NB methods tend to result in a substantially lower AUC with a higher standard error compared to our ANB methods due to the inclusion of non-informative gene-sets. The IBS kernel generally performs well, resulting in similar performances as the linear kernel when the effects are linear; and better performances than the linear kernel when the effects are non-linear. In particular, for non-linear effects settings, the IBS kernel leads to higher AUCs for our ANB procedure with smaller standard errors than the linear kernel. Our methods with PCA perform very similarly to methods without PCA in terms of prediction with very slight improvement in model size and prediction accuracy, but the computational efficiency is much greater when using PCA. Decreasing $\wp$ from 0.999 to 0.99 gives nearly identical results so we report only $\wp = 0.999$ which is approximately $1 - 1/n_t$. Overall, we observe the strengths of the PCA and adaptively weighted blocking models, and note that we obtain the best prediction accuracy with k_{IBS}. The average number of PCs included in the first stage for $\wp = 0.999$ (mean 26, median 12) is typically quite larger than those for $\wp = 0.99$ (mean 12, median 7). It is important to note that both procedures select substantially fewer PCs than the total number of nonzero eigenvalues, which is the number used in the noPCA

methods. Furthermore, most computational algorithms to estimate eigenvalues have difficulty exactly estimating true zero eigenvalues and so selecting all of the PCs corresponding to estimated nonzero eigenvalues can lead to much instability and can increase the computational burden, especially with large n. Thus, in general, we recommend $\wp = 0.999$ to achieve an excellent balance between prediction accuracy and computational ease.

Table 2.2: AUC $\times 100$ (empirical standard error) for the simulation studies. The columns represent the types of functions used to generate the outcome.

$\mathcal{K}$	$\wp$	Block Weighting	allL	LNL	NL
IBS	.999	ANB	80.9 (2.1)	87.5 (1.8)	84.1 (2.2)
LIN	.999	ANB	81.2 (2.3)	81.3 (2.5)	76.2 (3.2)
IBS	.999	NB	71.2 (2.3)	73.0 (2.6)	70.8 (2.8)
LIN	.999	NB	71.9 (2.4)	70.3 (2.6)	67.2 (2.9)
IBS	noPCA	ANB	80.7 (2.2)	87.4 (1.8)	84.0 (2.1)
LIN	noPCA	ANB	80.9 (2.4)	81.2 (2.5)	75.5 (3.5)
IBS	noPCA	NB	71.2 (2.3)	73.0 (2.5)	70.6 (2.8)
LIN	noPCA	NB	68.6 (5.9)	68.6 (2.8)	65.2 (3.1)
Global Method			allL	LNL	NL
gRidge			70.7 (2.4)	69.5 (2.5)	64.1 (2.9)
gKernelIBS			73.6 (2.3)	75.6 (2.3)	68.5 (2.8)
SIS			75.4 (3.9)	72.5 (5.0)	66.7 (4.2)
WLGR			65.7 (2.6)	63.8 (2.8)	57.5 (3.2)

We also conducted simulation studies to examine the effects of data partitioning on our estimates of ψ and γ. We proposed a cross-validation procedure on the training set to estimate γ in section 2.3.3. However, there are two simpler approaches to consider: (1) using the entire training set to estimate $h^{(m)}$ and the same entire training set to estimate γ, (2) dividing the training set into two non-overlapping parts, estimating $h^{(m)}$ on the first part and γ on the second part. As expected, our simulation studies revealed that method (1) led to overfitting in our estimates, and

method (2) was underpowered to accurately estimate $\boldsymbol{\psi}^{(m)}$ and $h^{(m)}$. The average AUCs from method (1) were consistently about 10% lower than the AUCs from our cross-validation method (γ-CV). The average AUCs from method (2) were more similar to AUCs from (γ-CV), but γ for gene-sets with weak effects were estimated as 0 much more often. Furthermore, when comparing AUCs of the naive bayes methods for (2) and (γ-CV), we saw that AUCs were lower for method (2), which implies that our estimates of $h^{(m)}$ with half of the training data in method (2) were less accurate than our estimates using the entire training set in (γ-CV). We conclude that using a cross-validated approach to estimate the gene-set weights improves estimation and prediction accuracy in smaller data sets where method (2) is underpowered. Additionally, this CV method avoids the overfitting that results from estimating parameters from both stages of estimation on the same training set.

2.4.3 Data example

Using the methods described above, we also constructed T1D risk prediction models based on the WTCCC GWAS dataset. To build our prediction model, we randomly selected 3500 subjects as a training set to implement the first stage and the cross-validation procedures for the second stage, and left the remaining 1500 subjects for a validation set.

Although our dataset includes 40 SNPs that are known to be associated with T1D disease status, they do not explain all of the genetic variability and there may be many other SNPs that are associated with disease status through complex effects. Furthermore, many autoimmune diseases may be caused by the same SNPs or genes and therefore investigating SNPs or genes associated with other autoimmune diseases

might improve prediction of T1D disease status. We hope to gain predictive power by allowing other SNPs to be included in the model via the gene-sets constructed with the NHGRI catalog.

We compare our methods to the same methods described in the simulation section. The AUC estimates in the validation set for selected procedures are shown in Table 2.3. In addition, we compare our methods to univariate SNP based methods that include only the 40 SNPs found to be associated with T1D in the model. We combine these 40 SNPs through either ridge regression, a kernel machine model with $\mathcal{K}_{\text{IBS}}$, and a weighted log odds ratio risk score (univariate WLGR). In general, our proposed ANB KM estimators have much higher AUC than the global methods and purely NB methods. Our prediction model obtains a high AUC ($\widehat{\text{AUC}} = 0.94$, median across 15 permutations of the data) via the ANB KM PCA method with k_{IBS} with $\wp = .999$. This method obtains almost identical results to the same method that does not use PCA, but it required much less computational time. Our procedure estimates 112 of the 350 gene-sets to have nonzero effects in the second stage. It includes 41 of the 92 genes that have been associated with T1D in the final model. The other 71 genes that were included in the model were not reported as being associated with T1D specifically, but have been shown to be linked to other autoimmune disease risk. The $\mathcal{K}_{\text{LIN}}$ ANB blockwise methods select many more gene-sets to be included in the final model and have much lower AUC.

The most common methods for risk prediction build a prediction rule based on univariate SNP testing and estimation. We see in this case that prediction accuracy can be greatly improved with our methods. By incorporating many more SNPs through a gene block structure and by adaptively weighting the blocks' effects we

can predict disease risk very accurately.

Table 2.3: AUC $\times 100$ for the models used to predict type 1 diabetes risk in the WTCCC dataset using 350 gene-sets. Median AUC across 15 random permutations of the dataset.

$\mathcal{K}$	$\wp$	Block Weighting	AUC	Model Size
IBS	.999	ANB	94.3	112
LIN	.999	ANB	84.5	344
IBS	.999	NB	85.5	350
LIN	.999	NB	83.6	350
IBS	noPCA	ANB	94.1	103
LIN	noPCA	ANB	85.1	344
IBS	noPCA	NB	84.4	350
LIN	noPCA	NB	82.2	350
Global Method			AUC	
gRidge			80.1	
gKernelIBS			82.2	
WLGR			82.0	
Ridge			76.1	
KernelIBS			78.1	
WLGR			78.3	

2.5 Discussion

The successful incorporation of genetic markers in risk prediction models has important implications in personalized medicine and disease prevention. However, standard methods for building such models are hindered by large datasets and nonlinear genetic associations with the outcome of interest. To overcome these challenges, we propose a multi-stage prediction model that includes genetic markers partitioned into gene-sets based on prior knowledge about the LD structure or pathway information. To achieve a good balance between allowing a flexible model that captures

complex interactive effects and efficient estimation in the model parameters, we utilize a NBKM regression framework that builds non-linear risk models separately for each gene-set and then aggregates information from multiple gene-sets efficiently via an adaptive block-wise weighting scheme. Through simulation studies and a real data example, we show that our NBKM model performs well and maintains high prediction accuracy even when the underlying association of covariates and disease risk is complex. We see that kernel PCA approximation improves over the noPCA methods mainly in the computational efficiency, but also slightly in model selection and prediction accuracy. Hence in practice, we would recommend applying the kernel PCA with a relatively stringent threshold such as $1 - n^{-1}$ although the optimal selection of threshold warrants further investigation.

Incorporating the block structure of the gene-sets in our model could potentially improve prediction accuracy over global methods that attempt to build one-stage models with a large number of unstructured genetic markers. Of course, one would expect that their relative performance may also depend on how well the gene-sets are grouped together. In our numerical studies, we partitioned the genome based on the gene structure. One may also consider forming sets of genetic markers based on other biological criteria such as linkage disequilibrium blocks or genetic pathways. We note that when partitioning the entire genome into gene-sets, one may first screen these blocks using a testing procedure such as the logistic kernel machine score test proposed by Liu et al. (2008) to reduce the number of blocks in the model which may improve efficiency and prediction accuracy. It would also be interesting to explore the best procedures for this initial screening stage. We have found the KM score test for associations within gene blocks to perform well in other numerical examples. However, further research is needed to explore how the proposed procedure is affected

by the screening procedure and the criteria used for forming the gene-sets.

Lastly, the proposed procedures can be easily extended to adjust for covariates. For example, if there are existing clinical variables or population structure principal components $\mathbf{X}$ available for risk prediction, one may impose a conditional NBKM model by extending (2.1) and (3.5) to

$$\begin{aligned} &\text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(\bullet)}, \mathbf{X}_i) = a_0 + \mathbf{X}_i^\intercal \mathbf{b}_0 + \sum_{m=1}^{M} \text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(l)}, \mathbf{X}_i) \\ \text{and} \quad &\text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(m)}, \mathbf{X}_i) = a_0^{(l)} + \mathbf{X}_i^\intercal \mathbf{b}_0^{(m)} + h^{(m)}(\mathbf{Z}_i^{(m)}), \quad \text{respectively.} \end{aligned}$$

The proposed procedures can be carried out by first fitting M separate models with $(\mathbf{X}_i, \mathbf{Z}_i^{(m)})$ and then adaptively weighting to obtain a sparse combination of $h^{(m)}$ across M gene-sets.

2.6 Appendix A: Proofs

2.6.1 Convergence of the kernel PCA estimator of $\widehat{h}^{(m)}(\mathbf{z}^{(m)})$

Here we show that our kernel PCA estimator $\widehat{h}^{(m)}(\mathbf{z}^{(m)}) = \widehat{\mathbf{\Psi}}_{(m)}(\mathbf{z}^{(m)})^\intercal \widehat{\boldsymbol{\beta}}^{(m)}$ from the NBKM is a root-n consistent estimator for $h^{(m)}(\mathbf{z}^{(m)})$ under the assumption that the dimension of the feature space $\mathcal{H}_k^{(m)}$ is finite. To simplify notation we drop some m subscripts and superscripts. By Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), $h(\mathbf{z}) \in \mathcal{H}_k$ has a *primal representation* $h(\mathbf{z}) = \sum_{j=1}^{r} \beta_j \psi_j(\mathbf{z})$, $\psi_j(\mathbf{z}) = \sqrt{\lambda_j}\phi_j(\mathbf{z})$, where $\{\lambda_j\}$ and $\{\phi_j\}$ are the eigenvalues and eigenfunctions of k under the probability measure $\mathcal{P}_{\mathbf{Z}}$, $r < \infty$ and $|\phi_j(\cdot)|$ is bounded by a constant C_φ.

First we show that asymptotically, the number of eigenvalues and eigenvectors

included in the kernel PCA model is greater or equal to the true number of nonzero eigenvalues. To this end, we recall that we select the r_n eigenvalues that satisfy $\sum_{j=1}^{r_n} l_j / \sum_{i=1}^{n} l_j \geq \wp = 1 - \epsilon \to 1$. Then

$$\begin{aligned} P(r_n < r) = P(r_n \leq r-1) &\leq P\left(\frac{\sum_{j=1}^{r-1} l_j}{\sum_{i=1}^{n} l_j} > 1 - \epsilon \right) && (2.16) \\ &= P\left(\epsilon < \frac{l_r + \sum_{j=r+1}^{n} l_j}{\sum_{i=1}^{n} l_j} \right) && (2.17) \\ &= o_p(n^{-1}) \to 0 && (2.18) \end{aligned}$$

since $l_r \to \lambda_r > 0$ and $\sum_{j=r+1}^{n} l_j / \sum_{j=1}^{n} l_j \to 0$ from Koltchinskii and Giné (2000, Theorem 3.1) and Braun (2005, Theorem 3.94). Therefore, $P(r_n \geq r) \to 1$. Hence in the sequel for the purpose of establishing $\mathrm{O}_P(n^{-\frac{1}{2}})$ convergence rate of $\widehat{h}(\mathbf{z}) - h(\mathbf{z})$ in probability, we only considers the realizations when $r_n \geq r$.

In order to obtain convergence of $\widehat{h}(\mathbf{z})$ we must have convergence of the eigensystem. We first address the asymptotic behavior of the eigenvectors and eigenfunctions in the following two subsections.

2.6.2 Convergence of eigenvectors within sample space

By the spectral theorem we may write the kernel function k as $k(s,t) = \sum_{j=1}^{r} \lambda_j \phi_j(s) \phi_j(t)$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_r$ and $\mathbb{E}\phi_j(X)\phi_l(X) = \delta_{jl}$, where δ_{ij} is the Kronecker's delta. As in the text, denote $\mathbf{K}_n$ the Gram matrix, i.e., $\mathbf{K}_n = n^{-1} \{k(\mathbf{Z}_i, \mathbf{Z}_j)\}_{1 \leq i,j \leq n}$. By these kernel properties and a singular value decomposition (SVD), we have that

$$\mathbf{K}_n = n^{-1} \boldsymbol{\Phi} \boldsymbol{\Lambda} \boldsymbol{\Phi}^{\intercal} = n^{-1} \widehat{\boldsymbol{\Phi}} \widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{\Phi}}^{\intercal}, \tag{2.19}$$

where $\mathbf{\Phi} = \{\phi_j(\mathbf{Z}_i)\}_{1\le i\le n, 1\le j,\le r}$, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, $\widehat{\mathbf{\Phi}} = [\mathbf{u}_1, ..., \mathbf{u}_n]$ and $\widehat{\mathbf{\Lambda}} = \mathrm{diag}(l_1, ..., l_n)$ are the eigenvector and eigenvalue matrices. It is clear that

$$\mathbb{E}\|n^{-1}\mathbf{\Phi}^\intercal\mathbf{\Phi} - I\|_F^2 = \mathbb{E}\sum_{j_1,j_2=1}^{r}\left(\mathbb{E}_n\phi_{j_1}(X)\phi_{j_2}(X) - \mathbb{E}\phi_{j_1}(X)\phi_{j_2}(X)\right)^2 \tag{2.20}$$

where $\mathbb{E}_n$ stands for the empirical expectation, i.e., $\mathbb{E}_n f(X) = n^{-1}\sum_{i=1}^n f(x_i)$, and $\|\cdot\|_F$ represents the Frobenius matrix norm. Letting the SVD of $\mathbf{\Phi}$ be $\mathcal{U}_{n\times r}\mathcal{D}_{r\times r}\mathcal{V}_{r\times r}^\intercal$, we have

$$\sum_j (n^{-1}d_j^2 - 1)^2 = \|n^{-1}\mathbf{\Phi}^\intercal\mathbf{\Phi} - I\|_F^2 = O_p(n^{-1}), \tag{2.21}$$

which implies that $\max_j |n^{-\frac{1}{2}}d_j - 1| = O_p(n^{-\frac{1}{2}})$. Write $\widetilde{\mathbf{G}} = \mathcal{U}\mathcal{V}^\intercal\mathbf{\Lambda}\mathcal{V}\mathcal{U}^\intercal$. Then

$$\begin{aligned}
\|\widetilde{\mathbf{G}} - \mathbf{K}_n\|_F^2 &= \|\mathcal{U}\mathcal{V}^\intercal\mathbf{\Lambda}\mathcal{V}\mathcal{U}^\intercal - n^{-1}\mathcal{U}\mathcal{D}\mathcal{V}^\intercal\mathbf{\Lambda}\mathcal{V}\mathcal{D}\mathcal{U}^\intercal\|_F^2 &(2.22)\\
&= \|\mathcal{V}^\intercal\mathbf{\Lambda}\mathcal{V} - n^{-1}\mathcal{D}\mathcal{V}^\intercal\mathbf{\Lambda}\mathcal{V}\mathcal{D}\|_F^2 &(2.23)\\
&\le \|\mathcal{V}^\intercal\mathbf{\Lambda}\mathcal{V}\|_{\max}^2\sum_{j_1,j_2=1}^{r}(1 - n^{-1}d_{j_1}d_{j_2}) &(2.24)\\
&= O_p(n^{-1}) &(2.25)
\end{aligned}$$

Recall that $\mathcal{U}\mathcal{V}^\intercal$ is the eigenvector of $\widetilde{\mathbf{G}}$ and $\mathbb{U}_{n\times r_n}\mathbb{D}_{r_n\times r_n}\mathbb{U}_{n\times r_n}^\intercal$ is the kernel PCA approximation to $\mathbf{K}_n$. From a standard perturbation analysis and the fact that $\sum_{j=r+1}^n l_j = O_p(n^{-\frac{1}{2}})$, there exists a constant $C > 0$ such that

$$\|n^{-\frac{1}{2}}\mathbb{U} - [\mathcal{U}\mathcal{V}^\intercal, \mathbf{0}_{n\times(r_n-r)}]\|_F^2 \le \|n^{-\frac{1}{2}}\widehat{\mathbf{\Phi}} - [\mathcal{U}\mathcal{V}^\intercal, \mathbf{0}_{n\times(n-r)}\|_F^2 \le C\|\widetilde{\mathbf{G}} - \mathbf{K}_n\|_F^2 = O_p(n^{-1}). \tag{2.26}$$

Finally, by the triangular inequality, $\widehat{\varepsilon}_U \equiv n^{-\frac{1}{2}}\|\mathbb{U} - [\mathbf{\Phi}, \mathbf{0}_{n\times(r_n-r)}]\|_F$ is bounded above by

$$\|n^{-\frac{1}{2}}\mathbf{\Phi} - \mathcal{U}\mathcal{V}^\intercal\|_F + \|n^{-\frac{1}{2}}\mathbb{U} - [\mathcal{U}\mathcal{V}^\intercal, \mathbf{0}_{n\times(r_n-r)}]\|_F = O_p(n^{-\frac{1}{2}}) \tag{2.27}$$

This, together with Theorem 3.1 of Koltchinskii and Giné (2000) regarding the convergence of the eigenvalues of $\mathbf{K}_n$, implies that

$$n^{-\frac{1}{2}}\|\widetilde{\mathbf{\Psi}} - [\mathbf{\Psi}, \mathbf{0}_{n\times(r_n-r)}]\|_F = O_p(n^{-\frac{1}{2}}). \tag{2.28}$$

2.6.3 Convergence of eigenvectors, Nyström projection

We now extend this result to include data points outside the sample space and show that the Nyström projection estimate of the eigenvector

$$\widehat{\mathbf{\Phi}}(\mathbf{z}) = \{\widehat{\phi}_j(\mathbf{z})\}_{1\le j\le r_n} = n^{-1}\mathbb{D}^{-1}_{r_n\times r_n}\mathbb{U}^{\intercal}_{n\times r_n}\mathbf{K}_{\mathbf{z}} \tag{2.29}$$

also converges to $[\mathbf{\Phi}(\mathbf{z}) = \{\phi_j(\mathbf{z})\}^{\intercal}_{1\le j\le r}, \mathbf{0}^{\intercal}_{r_n-r}]^{\intercal}$ at a root-n rate, where $\mathbf{K}_{\mathbf{z}} = [k(\mathbf{z}, \mathbf{Z}_1), ..., k(\mathbf{z}, \mathbf{Z}_n)]^{\intercal}$. To this end, recall that $\mathbf{\Phi} = \{\mathbf{\Phi}(\mathbf{Z}_i)\}_{1\le i\le n}$ and let $\mathbb{D}_{\le r} = \text{diag}\{l_1, ..., l_r\}$ and $\mathbb{U}_{\le r} = [\mathbf{u}_1, ..., \mathbf{u}_r]$. Then

$$\begin{aligned}
& \|\widehat{\mathbf{\Phi}}(\mathbf{z}) - [\mathbf{\Phi}(\mathbf{z})^{\intercal}, \mathbf{0}^{\intercal}_{r_n-r}]^{\intercal}\|_2 = \|\widehat{\mathbf{\Phi}}_{\le r}(\mathbf{z}) - \mathbf{\Phi}(\mathbf{z})\|_2 + O_p(n^{-\frac{1}{2}}) && (2.30)\\
= \; & \|n^{-1}\mathbb{D}^{-1}_{\le r}\mathbb{U}^{\intercal}_{\le r}\mathbf{K}_{\mathbf{z}} - \mathbf{\Phi}(\mathbf{z})\|_2 + O_p(n^{-\frac{1}{2}}) && (2.31)\\
\le \; & \|n^{-1}\mathbb{D}^{-1}_{\le r}(\mathbb{U}_{\le r} - \mathbf{\Phi})^{\intercal}\mathbf{K}_{\mathbf{z}}\|_2 + \|n^{-1}\mathbb{D}^{-1}_{\le r}\mathbf{\Phi}^{\intercal}\mathbf{K}_{\mathbf{z}} - \mathbf{\Phi}(\mathbf{z})\|_2 + O_p(n^{-\frac{1}{2}}) && (2.32)\\
\le \; & rl_r^{-1}\widehat{\varepsilon}_U \left\|n^{-\frac{1}{2}}\mathbf{K}_{\mathbf{z}}\right\|_2 + \|n^{-1}\mathbb{D}^{-1}_{\le r}\mathbf{\Phi}^{\intercal}\mathbf{K}_{\mathbf{z}} - \mathbf{\Phi}(\mathbf{z})\|_2 + O_p(n^{-\frac{1}{2}}) && (2.33)\\
= \; & rl_r^{-1}\widehat{\varepsilon}_U \left\|n^{-\frac{1}{2}}\mathbf{K}_{\mathbf{z}}\right\|_2 + \|n^{-1}\mathbb{D}^{-1}_{\le r}\mathbf{\Phi}^{\intercal}\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}(\mathbf{z}) - \mathbf{\Phi}(\mathbf{z})\|_2 + O_p(n^{-\frac{1}{2}}) && (2.34)\\
\le \; & rl_r^{-1}\widehat{\varepsilon}_U \left\|n^{-\frac{1}{2}}\mathbf{K}_{\mathbf{z}}\right\|_2 + C_{\varphi}\|n^{-1}\mathbb{D}^{-1}_{\le r}\mathbf{\Phi}^{\intercal}\mathbf{\Phi}\mathbf{\Lambda} - I\|_F + O_p(n^{-\frac{1}{2}}) && (2.35)\\
\le \; & rl_r^{-1}\widehat{\varepsilon}_U \left\|n^{-\frac{1}{2}}\mathbf{K}_{\mathbf{z}}\right\|_2 + C_{\varphi}\left\{\|n^{-1}\mathbf{\Phi}^{\intercal}\mathbf{\Phi}\|_F\|n^{-1}\mathbb{D}^{-1}_{\le r}\mathbf{\Lambda} - I\|_F + \|n^{-1}\mathbf{\Phi}^{\intercal}\mathbf{\Phi} - I\|_F\right\} \\
& + O_p(n^{-\frac{1}{2}}) && (2.36)
\end{aligned}$$

which is of order $O_p(n^{-\frac{1}{2}})$. The above result follows from $l_j \to \lambda_j > 0$ for $1 \le j \le r$ and $\|l_j - \lambda_j\| = O_p(n^{-\frac{1}{2}})$ (Koltchinskii and Giné, 2000, Theorem 3.1) .

Once again, now that we have convergence of the eigenvectors, it is clear that we obtain convergence of the estimate $\widehat{\mathbf{\Psi}}(\mathbf{z}) = \left\{\sqrt{l_j}\widehat{\phi}_j(\mathbf{z})\right\}_{1\leq j\leq r}$ to $\mathbf{\Psi}(\mathbf{z}) = \left\{\sqrt{\lambda_j}\phi_j(\mathbf{z})\right\}_{1\leq j\leq r}$ by Theorem 3.1 of Koltchinskii and Giné (2000) which proves the eigenvalues of $\mathbf{K}_n$ converge to the true eigenvalues almost surely, so we have

$$\|\widehat{\mathbf{\Psi}}(\mathbf{z}) - [\mathbf{\Psi}(\mathbf{z})^\intercal, \mathbf{0}_{r_n-r}^\intercal]^\intercal\|_2 = O_p(n^{-\frac{1}{2}}). \tag{2.37}$$

2.6.4 Convergence of $\widehat{h}(\mathbf{z})$

The convergence of $\widehat{h}(\mathbf{z})$ follows from the convergence of our eigenfunctions above. First, consider the case when $\mathbf{z}$ is a sample point $\mathbf{Z}_i$ and we are interested in $\widetilde{h}(\mathbf{Z}_i) - h(\mathbf{Z}_i)$ for $\widetilde{h}(\mathbf{Z}_i) = \widetilde{\boldsymbol{\psi}}_i\widehat{\boldsymbol{\beta}} = \sum_{j=1}^{r_n}\widetilde{\psi}_j(\mathbf{Z}_i)\widehat{\beta}_j$ where $\widetilde{\psi}_j(\mathbf{Z}_i) = u_{ji}\sqrt{l_j}$. By (2.28), $\widetilde{\psi}_j(\mathbf{Z}_i) = \sqrt{l_j}u_{ji} = \mathrm{O}_P(n^{-\frac{1}{2}})$. We have

$$n^{\frac{1}{2}}\left[\widetilde{h}(\mathbf{Z}_i) - h(\mathbf{Z}_i)\right] = \underbrace{n^{\frac{1}{2}}\sum_{j=1}^{r_n}\widetilde{\psi}_j(\widehat{\beta}_j - \beta_{j0})}_{A} + \underbrace{n^{\frac{1}{2}}\sum_{j=1}^{r_n}(\widetilde{\psi}_j - \psi_j)\beta_{j0}}_{B} \tag{2.38}$$

For A, we wish to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|n^{\frac{1}{2}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|\geq C}\left[\mathcal{L}^{(\mathrm{P})}(a,\boldsymbol{\beta};\widetilde{\mathbf{\Psi}}) - \mathcal{L}^{(\mathrm{P})}(a,\boldsymbol{\beta}_0;\widetilde{\mathbf{\Psi}})\right] < 0\right\} \geq 1-\epsilon \tag{2.39}$$

Let $\mathbf{q} = n^{\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, and for a matrix (or vector) $\mathbb{A}$ with r_n columns (elements), let $\mathbb{A}_{\leq r}$ and $\mathbb{A}_{>r}$ be the first r and remaining $r_n - r$ columns (elements). From Appendix 2.6.3, we have $\|\widetilde{\mathbf{\Psi}}_{\leq r} - \mathbf{\Psi}\|_F + \|\widetilde{\mathbf{\Psi}}_{>r}\|_F = O_p(n^{-\frac{1}{2}})$ and $\boldsymbol{\beta}_{0,>r} = \mathbf{0}$. This, together with a taylor series expansion, a law of large number, and a central limit theorem,

implies that

$$D_n(\mathbf{q}) \equiv \mathcal{L}^{(\mathrm{P})}(a, \boldsymbol{\beta}_0 + n^{-\frac{1}{2}}\mathbf{q}; \widetilde{\boldsymbol{\Psi}}) - \mathcal{L}^{(\mathrm{P})}(a, \boldsymbol{\beta}_0; \widetilde{\boldsymbol{\Psi}}) \tag{2.40}$$

$$= n^{-\frac{1}{2}}\left[\mathbf{Y} - g(a + \widetilde{\boldsymbol{\Psi}}_{\leq r}\boldsymbol{\beta}_{0,\leq r})\right]^{\mathsf{T}}\widetilde{\boldsymbol{\Psi}}\mathbf{q} - \frac{1}{2}\mathbf{q}^{\mathsf{T}}\left[n^{-1}\sum_{i=1}^{n}\xi(a + \widetilde{\boldsymbol{\psi}}_i^{\mathsf{T}}\boldsymbol{\beta}_0)\widetilde{\boldsymbol{\psi}}_i\widetilde{\boldsymbol{\psi}}_i^{\mathsf{T}}\right]\mathbf{q}$$

$$-2n^{-\frac{1}{2}}\tau\sum_{j=1}^{r}\beta_{0j}\mathbf{q} + n^{-1}\tau r\mathbf{q}^{\mathsf{T}}\mathbf{q} + o_P\left(n^{-1}\|\mathbf{q}\|^2\right) \tag{2.41}$$

$$= n^{-\frac{1}{2}}\left\{\mathbf{Y} - g(a + \boldsymbol{\Psi}\boldsymbol{\beta}_{0,\leq r})\right\}^{\mathsf{T}}\boldsymbol{\Psi}\mathbf{q}_{\leq r} - \frac{1}{2}\mathbf{q}_{\leq r}^{\mathsf{T}}\left\{n^{-1}\boldsymbol{\Psi}^{\mathsf{T}}\mathrm{diag}\{\xi(a + \boldsymbol{\Psi}\boldsymbol{\beta}_{0,\leq r})\}\boldsymbol{\Psi}\right\}\mathbf{q}_{\leq r}$$

$$-2n^{-\frac{1}{2}}\tau\sum_{j=1}^{r}\beta_{0j}\mathbf{q} + n^{-1}\tau r\mathbf{q}^{\mathsf{T}}\mathbf{q} + O_P(n^{-\frac{1}{2}})\|\mathbf{q}\| - O_P(n^{-\frac{1}{2}})\|\mathbf{q}\|^2 \tag{2.42}$$

$$\leq O_P(1)\|\mathbf{q}\| - \mathbf{q}_{\leq r}^{\mathsf{T}}\mathbb{A}\mathbf{q}_{\leq r} + O_P(n^{-\frac{1}{2}})\|\mathbf{q}\|^2 \tag{2.43}$$

where $\ell(y, s)$ denotes the log likelihood function for the logistic regression model $P(Y = y \mid s) = g(s)$ with outcome y and score s, $\xi(x) = g(x)\{1 - g(x)\}$, $n^{-\frac{1}{2}}\tau = o_p(1)$, and $\mathbb{A}$ is the limit of $n^{-1}\boldsymbol{\Psi}^{\mathsf{T}}\mathrm{diag}\{\xi(a + \boldsymbol{\Psi}\boldsymbol{\beta}_{0,\leq r})\}\boldsymbol{\Psi}$. Since $\mathbb{A}$ is positive definite with smallest eigenvalue $\iota_{\mathbb{A}}$ bounded away from 0, $-\mathbf{q}_{\leq r}^{\mathsf{T}}\mathbb{A}\mathbf{q}_{\leq r} \leq -\iota_{\mathbb{A}}\|\mathbf{q}_{\leq r}\|_2^2$. When $\|\mathbf{q}\| \geq C$, we select a C sufficiently large so that the second term in (2.43) dominates D_n and so

$$P\left\{\sup_{\|\mathbf{q}\|\geq C}\left[\mathcal{L}^{(\mathrm{P})}(a, \boldsymbol{\beta}_0 + n^{-\frac{1}{2}}\mathbf{q}, \mathbf{0}; \widetilde{\boldsymbol{\Psi}}) - \mathcal{L}^{(\mathrm{P})}(a, \boldsymbol{\beta}_0; \widetilde{\boldsymbol{\Psi}})\right] < 0\right\} \to 1. \tag{2.44}$$

Therefore, $n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = O_p(1)$ and $n^{\frac{1}{2}}\sum_{j=1}^{r_n}\widetilde{\psi}_j(\widehat{\beta}_j - \beta_{j0}) = O_P(1)$.

For B, $\sum_{j=1}^{r} n^{\frac{1}{2}}\left(\widetilde{\psi}_j - \psi_j\right)\beta_{j0} = O_P(1)$ by (2.28). Using similar techniques with the Nyström projection estimate $\widehat{\boldsymbol{\Psi}}$ and the result (2.37) it follows that $\widehat{h}(\mathbf{z}) = \widehat{\boldsymbol{\Psi}}(\mathbf{z})\widehat{\boldsymbol{\beta}} = \sum_{j=1}^{r_n}\sqrt{l_j}\widetilde{\phi}_j(\mathbf{z})\widehat{\beta}_j$ also converges to $h(\mathbf{z})$ at a $n^{\frac{1}{2}}$ rate for $\mathbf{z}$ outside the sample space. Therefore, taking $\widehat{h}(\mathbf{z})$ to be $\widetilde{h}(\mathbf{z})$ for $\mathbf{z}$ in the sample, we have

$$n^{\frac{1}{2}}\left[\widehat{h}(\mathbf{z}) - h(\mathbf{z})\right] = O_P(1) \tag{2.45}$$

for all $\mathbf{z}$.

2.6.5 The oracle of gene-set weights

Here we prove that the regularized estimate $\widehat{\boldsymbol{\gamma}}$ exhibits the oracle sparsity property, in that $\lim_{n\to\infty}\mathcal{P}(\widehat{\boldsymbol{\gamma}}_{\mathcal{A}^C} = \mathbf{0}) = 1$, where $\widehat{\boldsymbol{\gamma}}$ is defined as in (3.18) and $\mathcal{A}^C = \{m : h^{(m)}(\mathbf{z}) = 0\}$ with its complement $\mathcal{A} = \{m : h^{(m)}(\mathbf{z}) \neq 0\}$. We may also reparameterize the objective function from (3.18) by defining $\theta_m = \gamma_m\|\widehat{h}^{(m)}\|$, $\|\widehat{h}^{(m)}\| = \sqrt{n^{-1}\sum_{i=1}^{n}(\widehat{h}^{(m)}(\mathbf{Z}_i^{(m)}))^2}$, $\widetilde{H}_{im} = \widehat{h}^{(m)}(\mathbf{Z}_i^{(m)})/\|\widehat{h}^{(m)}\|$, $\widetilde{\mathbf{H}}_m = [\widetilde{H}_{1m}, ..., \widetilde{H}_{nm}]^\intercal$, and $\widetilde{\mathbb{H}} = [\widetilde{\mathbf{H}}_1, ..., \widetilde{\mathbf{H}}_M]$. The reparameterized estimator can then be represented as

$$\{\hat{b}, \widehat{\boldsymbol{\theta}}\} = \underset{b,\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}_{\widetilde{\mathbb{H}}}(b, \boldsymbol{\theta}) - \tau_2 \sum_{m=1}^{M} \frac{|\theta_m|}{\|\widehat{h}^{(m)}\|} \tag{2.46}$$

where

$$\mathcal{L}_{\widetilde{\mathbb{H}}}(b, \boldsymbol{\theta}) = \mathbf{Y}^\intercal \log g(b + \widetilde{\mathbb{H}}\boldsymbol{\theta}) + (1 - \mathbf{Y})^\intercal \log\{1 - g(b + \widetilde{\mathbb{H}}\boldsymbol{\theta})\}. \tag{2.47}$$

Now let $\mathbf{u} = n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and define $\phi(\widetilde{\mathbb{H}}^\intercal\boldsymbol{\theta}) = \log(1 + \exp(\widetilde{\mathbb{H}}^\intercal\boldsymbol{\theta}))$. Then we have

$$\begin{aligned} V_n(\mathbf{u}) \;\equiv\; & \left[\mathcal{L}_{\widetilde{\mathbb{H}}}(b, \boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}) - \tau_2 \sum_{m=1}^{M} \frac{|\theta_{0m} + n^{-\frac{1}{2}}u_m|}{\|\widehat{h}^{(m)}\|}\right] \\ & - \left[\mathcal{L}_{\widetilde{\mathbb{H}}}(b, \boldsymbol{\theta}_0) - \tau_2 \sum_{m=1}^{M} \frac{|\theta_{0m}|}{\|\widehat{h}^{(m)}\|}\right] \qquad (2.48) \\ =\; & n^{-\frac{1}{2}}\left[\mathbf{Y} - \phi'(b + \widetilde{\mathbb{H}}^\intercal\boldsymbol{\theta}_0)\right]\widetilde{\mathbb{H}}^\intercal\mathbf{u} - (2n)^{-1}\mathbf{u}^\intercal\left[\widetilde{\mathbb{H}}^\intercal\phi''(b + \widetilde{\mathbb{H}}^\intercal\boldsymbol{\theta}_0)\widetilde{\mathbb{H}}\right]\mathbf{u} \\ & - n^{-\frac{1}{2}}\tau_2 \sum_{m=1}^{M} \frac{n^{\frac{1}{2}}\left[|\theta_{0m} + n^{-\frac{1}{2}}u_m| - |\theta_{0m}|\right]}{\|\widehat{h}^{(m)}\|} + o_p(n\|\mathbf{u}\|^2) \qquad (2.49) \end{aligned}$$

We know the first two components of (2.49) are $O_p(1)$, converging to some function $V(\mathbf{u})$, so consider the limiting distribution of the third term. When $m \in \mathcal{A}$,

$n^{-\frac{1}{2}}\left[|\theta_{0m}+n^{-\frac{1}{2}}u_m|-|\theta_{0m}|\right] \rightarrow u_m \text{sgn}(\theta_{0m}), n^{-\frac{1}{2}}\tau_2 \rightarrow 0, \|\widehat{h}^{(m)}\| \rightarrow C > 0$. Therefore, by Slutsky's theroem, the third term converges to 0. However, when $m \in \mathcal{A}^C$, $n^{\frac{1}{2}}(|\theta_{0m}+n^{-\frac{1}{2}}u_m|-|\theta_{0m}|) = |u_m|, \tau_2 \rightarrow \infty$, $\|\widehat{h}^{(m)}\| = O_p(n^{-\frac{1}{2}})$. Thus, we have

$$V_n(\mathbf{u}) \rightarrow \begin{cases} V(\mathbf{u}) & \text{if } u_m = \mathbf{0}, \forall m \in \mathcal{A}^C \\ \infty & \text{if } u_m \neq \mathbf{0}, \forall m \in \mathcal{A}^C \end{cases} \tag{2.50}$$

Since V_n is convex in $\mathbf{u}$ and V has a unique minimum, we can use the epi-convergence results (Geyer, 1994) as in Knight and Fu (2000) and Zou (2006) to show that $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{0\mathcal{A}}) = O_p(1)$ and $n^{\frac{1}{2}}\widehat{\boldsymbol{\theta}}_{\mathcal{A}^C} \rightarrow_d 0$.

Now we show that for all $m \in \mathcal{A}^C, P(\widehat{\theta}_m \neq 0) \rightarrow 0$. For the event $\widehat{\theta}_m \neq 0$, the Karush-Kuhn-Tucker (KKT) optimality conditions imply that

$$\widetilde{\mathbf{H}}_m^{\intercal}\left[\mathbf{Y} - \phi'(b + \widetilde{\mathbb{H}}\widehat{\boldsymbol{\theta}})\right] = \frac{\tau_2}{\|\widehat{h}^{(m)}\|} \tag{2.51}$$

From the taylor expansion along with similar arguments as in Zou (2006) and as above, we have

$$\begin{aligned} n^{-\frac{1}{2}}\widetilde{\mathbf{H}}_m^{\intercal}\left[\mathbf{Y} - \phi'(b + \widetilde{\mathbb{H}}\widehat{\boldsymbol{\theta}})\right] &= n^{-\frac{1}{2}}\widetilde{\mathbf{H}}_m^{\intercal}\left[\mathbf{Y} - \phi'(b + \widetilde{\mathbb{H}}\boldsymbol{\theta}_0)\right] + \\ &\quad n^{-1}\widetilde{\mathbf{H}}_m^{\intercal}\text{diag}\{\phi''(b + \widetilde{\mathbb{H}}\boldsymbol{\theta}_0)\}\widetilde{\mathbb{H}}\left[n^{\frac{1}{2}}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})\right] + o_p(1) \\ &= O_p(1). \end{aligned} \tag{2.52}$$

However, $n^{\frac{1}{2}}\tau_2 \rightarrow \infty$ while $n^{\frac{1}{2}}\|\widehat{h}^{(m)}\| = O_p(1)$ so $\frac{\tau_2}{\|\widehat{h}^{(m)}\|} \rightarrow \infty$. Therefore,

$$P(\widehat{\theta}_m \neq 0) \leq P\left\{\widetilde{\mathbf{H}}_m^{\intercal}\left[\mathbf{Y} - \phi'(b + \widetilde{\mathbb{H}}\widehat{\boldsymbol{\theta}})\right] = \frac{\tau_2}{\|\widehat{h}^{(m)}\|}\right\} \rightarrow 0. \tag{2.53}$$

Therefore, since $\widehat{\gamma}_m = 0$ implies $\widehat{\theta}_m = 0$ almost everywhere, we have proved that $\widehat{\boldsymbol{\gamma}}$ exhibits the oracle property.

2.7 Appendix B. Simulation details

For the simulation settings, we generated disease status through various functions of the SNPs in four regions. Specifically, $\text{logitpr}(Y = 1 \mid \mathbf{Z}^{(\bullet)}) = \sum_{m=1}^{4} h^{(m)}(\mathbf{z})$, where $h^{(m)}(\mathbf{z}) = h^{(NLm)}(\mathbf{z})$ for the nonlinear (allNL) model, $h^{(m)}(\mathbf{z}) = h^{(Lm)}(\mathbf{z})$ for the linear (allL) model, and $h^{(m)}(\mathbf{z}) = h^{(Lm)}(\mathbf{z}), m = 1, 2$ and $h^{(m)}(\mathbf{z}) = h^{(NL(m-1))}(\mathbf{z}), m = 3, 4$ for the partially linear and nonlinear (NLN) model. The forms of these functions are as follows: $h^{(NL1)}$ includes many two- and three-way interactions, $h^{(NL2)}$ has exponential effects and a many-way interation, $h^{(NL3)}$ includes exponential effects, a many-way interaction, and a $\tan(\mathbf{z})$ function with 24 causal SNPs, $h^{(NL4)}$ includes exponential effects and a $\sin(\mathbf{z})$ function with 10 causal SNPs, $h^{(L1)}$ is additive for ten SNPs with equal weight, $h^{(L2)}$ is additive for all SNPs in a region with equal weight, $h^{(L3)}$ is additive for 12 SNPs with 6 having a small weight of .1 and the others a weight of .6, $h^{(L4)}$ is additive of 10 third of the SNPs in a region with 5 having a small weight of .35 and the others having a weight of .75.

Genetic risk classification via kernel machine methods for meta-analyses with heterogenous sampling schemes

Jessica Minnier[1], Eli Stahl[2], Robert Plenge[2], and Tianxi Cai[1]

[1]Department of Biostatistics
Harvard School of Public Health

[2]Division of Genetics
Division of Rheumatology, Immunology and Allergy
Brigham and Women's Hospital
Harvard Medical School

3.1 Introduction

Analyses of genetic association studies have provided major insights into the architecture of complex human disease, often focusing on strong associations of single markers or genes. Clinicians and statisticians alike strive to elucidate the effects of genetic patterns on disease risk and development. The discovery of many single nucleotide polymorphisms (SNPs) and other genetic markers that are highly associated with disease reveal the potential to accurately predict disease outcomes based on a patient's genetic profile. However, building accurate prediction models to classify disease risk remains difficult and can not solely rely on results from testing and estimation procedures.

The standard approach for harnessing information from a genetic study involves univariate testing of the association of individual markers with the outcome of interest. These univariate tests seek to identify the most highly associated markers that reach a stringent level of significance in the data. Often, different studies will result in discrepant results (Ioannidis et al., 2001, 2003). This can be a result of false positives and negatives, as well as a variety of differences in the datasets, such as heterogenous populations and sampling schemes. Replication of results is further hindered by underpowered studies that test association of thousands or millions markers with data from a much smaller number of subjects. Therefore, markers with weak effects or even rare alleles with strong effects can remain hidden. Generating larger data sets is always ideal, but can be prohibitively expensive and time consuming, especially for rare diseases. As a result, methods for meta-analyses have been adapted from the clinical trial setting to be utilized for combining data from multiple genetic studies to increase power in detecting associations.

Association results from meta-analyses have improved upon results from smaller individual studies by identifying alleles with more moderate effect sizes (Lohmueller et al., 2003; Zeggini et al., 2008; Stahl et al., 2010). For most complex diseases, however, only a fraction of the estimated genetic variation, or heritability, is explained by the implicated markers. There has been much speculation regarding the reasons for this so-called "missing heritability," including arguments that studies have still bee too underpowered to identify many common alleles with weak effects or rare alleles with stronger effects (Maher, 2008; Manolio et al., 2009; Eichler et al., 2010; Gibson, 2012). Furthermore, it is likely that complex diseases are influenced by complex effects of SNPs and genes that are not accurately represented by standard methods that assume markers are additively associated with disease.

A common approach to incorporate genotype information into risk prediction is to construct a genetic score from the total number of risk alleles or sum of log expression levels. Such a genetic score is then included as a new variable in the risk prediction model and assessed for its incremental value in risk prediction. However, supporting the missing heritability perception, little improvement in prediction accuracy is gained by adding such simple risk scores to the prediction model based on clinical markers (Gail, 2008; Meigs et al., 2008; Lee et al., 2012). Polygenic methods have been developed with the aim of predicting disease risk with a risk score incorporating a much larger number of markers (Evans et al., 2009; Purcell et al., 2009), but again, these additive burden models often fail to explain much more heritability than has been already accounted for with clinical markers and published risk alleles (Makowsky et al., 2011; Machiela et al., 2011).

In this paper, we address two main challenges of genetic risk prediction by con-

structing a prediction model that (1) incorporates complex effects of a large number of markers, and (2) combines data across heterogenous studies to increase prediction accuracy for future patients. Explicitly modeling complex effects such as interactions and nonlinear effects is a difficult problem that requires prior knowledge of the true genetic architecture, as well as large sample sizes. Therefore, we implement the Adaptive Naive Bayes Kernel Machine (ANBKM) proposed in Chapter 2 that efficiently captures nonlinear effects a via a kernel machine (KM) PCA regression framework without specifying a functional form (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002). This multi-stage method also leverages the biological structure of the genome by building the prediction model within gene-sets or regions and then aggregating over gene-set effects. Additionally, the ANBKM allows for incorporation of clinical covariates and population stratification variables (Price et al., 2006) that can account for differences in study populations. We extend these methods to applications for meta-analyses by developing a weighted version of this method that incorporates weights based on case-control sampling probabilities. This allows for accurate prediction even when combining data from studies with heterogeneous sampling schemes, even complex nested case-control studies. The weights are derived from the density-weighted Nyström method developed by Zhang and Kwok (2009). The Nyström method estimates the eigenfunctions of prohibitively large kernel matrix $\mathbf{K}$ with the eigenfunctions of a smaller $\widetilde{\mathbf{K}}$ constructed from a randomly sampled subset of the data (Williams and Seeger, 2001). We implement the density-weighted version of this method to estimate the eigenfunctions of the true kernel matrix of the entire unobservable population from which we sample our genetic data.

The structure of this paper is as follows. We first describe the established kernel machine methods for prediction in section 3.2. In section 3.3 we present our methods

for a weighted representation of the ANBKM model and in section 3.4 we describe procedures for implementing these methods in meta-analyses. In section 3.5 we summarize the complete meta-analysis procedure. We then illustrate the practical use of our methods by detailing a case study where we classify rheumatoid arthritis (RA) risk with data from 6 genome wide association studies (GWAS) in section 3.6. In section 3.7 we summarize simulation studies. We close with discussion and remarks in section 3.8.

3.2 Kernel Machine Methods

3.2.1 Logistic Kernel Machine Regression

First consider a single general data set consisting of n subjects with the genetic predictors $\mathbf{z}$ of subject i denoted $\mathbf{Z}_i$, and the binary outcome of interest as Y_i, with $Y_i = 1$ being diseased and $Y_i = 0$ being non-diseased. We may assume that Y is related to $\mathbf{z}$ through an unknown centered smooth function $h(\cdot)$ so that

$$\text{logitpr}(Y \mid \mathbf{z}) = a + h(\mathbf{z}). \tag{3.1}$$

The logistic kernel machine (KM) model (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002) assumes $h(\cdot) \in \mathcal{H}_k$ where $\mathcal{H}_k$ is the functional space implicitly specified by a positive definite kernel function $k(\cdot, \cdot)$ that measures the similarity or distance between a pair of genetic markers. The functional space $\mathcal{H}_k$ is spanned by a set of basis vectors $\psi_j(\mathbf{z}) = \sqrt{\lambda_j}\phi_j(\mathbf{z})$, such that $\{\lambda_j\}$ and $\{\phi_j\}$ are the eigenvalues and eigenfunctions of k under the probability measure $\mathcal{P}_{\mathbf{z}}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $\mathcal{P}_{\mathbf{z}}$ is the distribution of $\mathbf{z}$. Therefore, by Mercer's Theorem

(Cristianini and Shawe-Taylor, 2000), any $h \in \mathcal{H}_k$ has a *primal representation*,

$$h(\mathbf{z}) = \sum_{j=1}^{\infty} \beta_j \psi_j(\mathbf{z}), \tag{3.2}$$

and by the representer theorem (Kimeldorf and Wahba, 1971), the maximum penalized likelihood estimator for $h \in \mathcal{H}_k$ must admit a *dual representation*,

$$h(\mathbf{z}) = \sum_{j=1}^{n} \alpha_i k(\mathbf{z}, \mathbf{Z}_i). \tag{3.3}$$

These results reveal that the kernel function k projects the covariates $\mathbf{z}$ into the feature space $\mathcal{H}_k$ where $h(\mathbf{z})$ is a linear combination of the underlying basis functions. The choice of kernel therefore determines the form of h in the original space and hence influences predictive performance of the model. Examples of kernels include the simple linear kernel $k_{\text{LIN}} = \mathbf{z}_1^\intercal \mathbf{z}_2$ that implicitly assumes the simple logistic regression model, and polynomial kernels $k_{\text{POLY}}(\mathbf{z}_1, \mathbf{z}_2; d) = (1 + \mathbf{z}_1^\intercal \mathbf{z}_2)^d$ corresponding to d-way multiplicative interactive effect. The Gaussian Kernel, $k_{\text{GAU}}(\mathbf{z}_1, \mathbf{z}_2) = \exp\{-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\rho\}$ with tuning parameter ρ, assumes the radial basis and allows for complex non-linear smooth effects. The IBS kernel is $k_{\text{IBS}}(\mathbf{z}_1, \mathbf{z}_2) = \sum_{l=1}^{p} \text{IBS}(z_{1l}, z_{2l})$, where $\text{IBS}(z_{1l}, z_{2l})$ represents the number of alleles shared identity by state. There are strong arguments in favor of using these nonlinear kernels to capture interactive effects of genetic markers (Kwee et al., 2008; Wu et al., 2010). There are many other possible kernel functions and much work has been done in exploring various types of kernels (Wessel and Schork, 2006; Lin and Schaid, 2009; Mukhopadhyay et al., 2010). The kernel should satisfy Mercer's theorem (Cristianini and Shawe-Taylor, 2000) and be selected *a priori* based on biological knowledge of the between Y and $\mathbf{Z}$.

3.2.2 Adaptive Naive-Bayes Kernel Machine (ANBKM) Classification Model

We are interested in the setting when we assume that the genetic predictors of subject i, denoted $\mathbf{Z}_i^{(\bullet)}$, can be divided into M non-overlapping gene-sets so that $\mathbf{Z}_i^{(\bullet)} = \{\mathbf{Z}_i^{(1)}, \ldots, \mathbf{Z}_i^{(M)}\}$. We may build a prediction model that incorporates possible complex associations of the genetic predictors with the outcome while leveraging the biological information in the gene-set structure by implementing a logistic Adaptive Naive Bayes Kernel Machine (ANBKM) model (propsed in Chapter 2). This model estimates $\text{pr}(Y_i = 1 \mid \mathbf{Z}_i^{(\bullet)})$ for subject i by aggregating over gene-set effects

$$\text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(\bullet)}) = a + \sum_{m=1}^{M} \gamma_m \text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(m)}), \tag{3.4}$$

where $\boldsymbol{\gamma}$ is estimated via the penalized regression procedure LASSO. Within the mth gene-set, $\text{pr}(Y_i = 1 \mid \mathbf{Z}_i^{(m)})$ is estimated with a logistic kernel machine model

$$\text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(m)}) = a^{(m)} + h^{(m)}(\mathbf{Z}_i^{(m)}), \tag{3.5}$$

where $h^{(m)}(\cdot) \in \mathcal{H}_k^{(m)}$ is an unknown centered smooth function and the functional space $\mathcal{H}_k^{(m)}$ is implicitly specified by the chosen kernel function $k(\cdot, \cdot)$ Therefore, as in (3.2) $h^{(m)} \in \mathcal{H}_k^{(m)}$ has a *primal representation*,

$$h^{(m)}(\mathbf{z}) = \textstyle\sum_{j=1}^{\infty} \beta_j^{(m)} \psi_j^{(m)}(\mathbf{z}), \tag{3.6}$$

As outlined in Chapter 2, since the basis functions $\{\psi_j^{(m)}(\mathbf{z})\}$ are intractable in general, we may estimate these basis functions with eigenvalues and eigenvectors obtained via kernel principal components analysis (PCA) on the gram matrix $\mathbf{K} = n^{-1}\{k(\mathbf{Z}_i, \mathbf{Z}_j)\}_{1 \le i \le n, 1 \le j \le n}$. To gain efficiency and stability we take advantage of the decay of eigenvalues of k and include only the leading $r_n^{(m)}$ eigenvalues and

eigenvectors in our model, where $r_n^{(m)}$ is the smallest r such that $\sum_{i=1}^{r} l_i^{(m)} / \sum_{i=1}^{n} l_i^{(m)}$ is greater than a large pre-specified proportion. We then estimate $\boldsymbol{\beta}$ through a logistic ridge regression model using the leading $r_n^{(m)}$ basis function estimates. We estimate $h(\mathbf{z}^{(\bullet)})$ for future subjects with genetic data $\mathbf{z}^{(\bullet)}$ with the Nyström method for approximating eigenfunctions (Rasmussen, 2004).

3.3 Weighted Estimation of Kernel Eigenfunctions

Our estimates of $h(\mathbf{z})$ depend heavily on our estimates of the basis functions of k. In kernel machine learning theory, it is common to define the eigenvalues and eigenfunctions of the kernel function k by relating them via the integral

$$\int k(\mathbf{z}', \mathbf{z}) \phi_j(\mathbf{z}) d\mathcal{P}_{\mathbf{z}}(\mathbf{z}) = \lambda_j \phi_j(\mathbf{z}'). \tag{3.7}$$

where $\mathcal{P}_{\mathbf{z}}$ is the underlying probability distribution of the covariate vector $\mathbf{z}$ (Williams and Seeger, 2001). Then, as a consequence of Mercer's Theorem, if the data $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ are drawn *i.i.d.* from $\mathcal{P}_{\mathbf{z}}$, we may numerically approximate the integral with

$$n^{-1} \sum_{i=1}^{n} k(\mathbf{z}', \mathbf{Z}_i) \phi_j(\mathbf{Z}_i) \approx \lambda_j \phi_j(\mathbf{z}'). \tag{3.8}$$

This equation is fundamental for justifying the use of the kernel PCA and the Nyström methods for estimating $\psi_j(\mathbf{z}) = \sqrt{\lambda_j} \phi_j(\mathbf{z})$. However, genetic studies often do not have an underlying *i.i.d* sampling scheme and so $\mathcal{P}_{\mathbf{z}}$ is not constant over the data $\{\mathbf{Z}_i\}_{1 \leq i \leq n}$. For example, case-control studies are one of the most common sampling schemes for constructing GWAS to study rare diseases. In this method, rare disease cases are oversampled from the underlying population so that the pro-

portion of cases in the observed data is much higher than the prevalence. Hence, $\widetilde{\mathcal{P}}_{\mathbf{z}}$, the distribution of the observed data, does not equal $\mathcal{P}_{\mathbf{z}}$, but instead depends on disease status of subject i. Each data point $\mathbf{Z}_i$ should not be given equal weight in the estimation of the basis functions of k. This could result in a form of sample selection bias in our estimates of ψ_j and hence bias in our estimates of $h(\mathbf{z})$ in our prediction model. In order to adjust for possible sample selection bias in our estimates, we propose to adapt the density-weighted Nyström method presented in Zhang and Kwok (2009). The Nyström method has been used to estimate the eigenfunctions of a prohibitively large matrix K with the eigenfunctions of a smaller $\widetilde{K}$ constructed from a randomly sampled subset of the data (Williams and Seeger, 2001). We utilize the density-weighted version of the Nyström method to estimate the eigenfunctions of the true kernel matrix of the entire unobservable population from which we sample our genetic data. Zhang and Kwok (2009) extended the Nyström method by weighting by $\widetilde{\mathcal{P}}_{\mathbf{z}}$ at each of the sampled points. For the purposes of adjusting for sampling schemes in our estimates, we may assume $\widetilde{\mathcal{P}}_{\mathbf{z}}(\mathbf{Z}_i) = w_i \mathcal{P}_{\mathbf{z}}(\mathbf{Z}_i)$, with $\sum_{i=1}^{n} w_i = 1$. For example, a standard case-control study with the prevalence of the disease population π and n_1 cases and n_0 cases would have weights $w_i = \pi/n_1$ if subject i is a case, and $w_i = (1-\pi)/n_0$ if subject i is a control. Therefore, the integral in (3.7) and numerical result of (3.8) may be approximated as

$$\lambda_j \widetilde{\phi}_j(\mathbf{z}') = \int k(\mathbf{z}', \mathbf{z}) \widetilde{\phi}_j(\mathbf{z}) d\widetilde{\mathcal{P}}_{\mathbf{z}}(\mathbf{z}) \approx n^{-1} \sum_{i=1}^{n} w_i k(\mathbf{z}', \mathbf{Z}_i) \widetilde{\phi}_j(\mathbf{Z}_i) \tag{3.9}$$

where $\widetilde{\phi}_j$ are the eigenfunctions of this weighted representation of k. We may approximate $\widetilde{\phi}_j$ with $\tilde{\mathbf{u}}$ via the eigenvalue decomposition $n^{-1}\mathbf{K}\mathbf{W}\tilde{\mathbf{u}}_j = \widetilde{l}_j \tilde{\mathbf{u}}_j$, where $\widetilde{l}_j$ and $\tilde{\mathbf{u}}_j$ are the eigenvalues and eigenvectors of the asymmetric matrix $\mathbf{K}\mathbf{W} = \mathbf{K}\text{diag}\{w_i\}_{1 \le i \le n} = \{w_i k(\mathbf{Z}_i, \mathbf{Z}_j)\}_{1 \le i \le n}$. Through the variable transforma-

tion $\mathbf{u}_j = \mathbf{W}^{1/2}\tilde{\mathbf{u}}_j$ and by defining $\widetilde{\mathbf{K}} = \mathbf{W}^{1/2}\mathbf{K}\mathbf{W}^{1/2}$ we can transform this equation into a symmetric eigenvalue problem so that $n^{-1}\widetilde{\mathbf{K}}\mathbf{u}_j = \widetilde{l}_j\mathbf{u}_j$. Therefore, our final estimate of the eigenfunction for a sample point is $\tilde{\mathbf{u}} = \mathbf{W}^{-1/2}\mathbf{u}$. Adapting the estimate of the set of basis functions $\boldsymbol{\Psi}_{(m)}(\mathbf{z}^{(m)}) = \{\psi_j(\mathbf{z}^{(m)})\}_{1\leq j\leq r_n}$ in Chapter 2, we have:

$$\widehat{\boldsymbol{\Psi}}_{(m)}(\mathbf{z}^{(m)}) = n^{-1}\left[w_1 k(\mathbf{z}^{(m)}, \mathbf{Z}_1^{(m)}), \ldots, w_n k(\mathbf{z}^{(m)}, \mathbf{Z}_n^{(m)})\right]\widetilde{\mathbb{U}}_{(m)}\text{diag}\left(\frac{1}{\sqrt{\widetilde{l}_1^{(m)}}}, \cdots, \frac{1}{\sqrt{\widetilde{l}_{r_n}^{(m)}}}\right) \tag{3.10}$$

where $\widetilde{\mathbb{U}}_{(m)} = \left[\tilde{\mathbf{u}}_1^{(m)}, \ldots, \tilde{\mathbf{u}}_{r_n}^{(m)}\right]$. Therefore, our estimate of h is

$$\widehat{h}^{(m)}(\mathbf{z}^{(m)}) = \widehat{\boldsymbol{\Psi}}_{(m)}(\mathbf{z}^{(m)})\widehat{\boldsymbol{\beta}}^{(m)} \tag{3.11}$$

3.4 Meta-Analysis Model

3.4.1 Notation

Heretofore we have merely been focusing on single datasets. When we introduce multiple data sets from independent sources with possibly heterogenous sampling schemes, we introduce more challenges in building a prediction model.

Now suppose we have a set of S data-sets, $\mathcal{S} = \{\mathcal{D}^{(s)} = \{Y_i, \mathbf{Z}_i^{(s)(\bullet)}\}_{1\leq i\leq n_s} | s = 1 \ldots S\}$, with each consisting of data from n_s subjects drawn from the general population via independent studies. We may divide the set of datasets into a set of "training sets" denoted $\mathcal{S}_{t1}$ that we use to estimate our kernel machine model parameters $\boldsymbol{\Psi}_{(m)}, \boldsymbol{\beta}_{(m)}$, a second set of "training sets" denoted $\mathcal{S}_{t2}$ that we use to estimate the gene-set weights $\boldsymbol{\gamma}$, and a set of "validation sets" denoted $\mathcal{S}_v$ that we use to estimate the prediction accuracy of our final model. Denote the genetic covariate data from

the n_s subjects in study s as $\{\mathbf{Z}_i^{(s)(\bullet)}\}_{1\leq i\leq n_s}$ so that the data in gene-set m for study s is denoted $\{\mathbf{Z}_i^{(s)(m)}\}_{1\leq i\leq n_s}$.

3.4.2 Combining Estimates of Eigenfunctions Across Studies

When estimating $h(\mathbf{z})$ with multiple datasets that were drawn from the population via heterogeneous sampling schemes, special care must be taken to estimate the eigenfunctions from each study accurately before combining estimates across studies. First, consider the eigenfunctions estimated using the data from using study s, $\mathcal{D}^{(s)} \in \mathcal{S}_{t1}$, as the training set. Implementing the weighted estimation procedures detailed in section 3.3, we obtain estimates of the eigenfunctions through Nyström projection on the entire meta-data:

$$\widehat{\mathbf{\Psi}}_{(m)}^{(s)}(\mathbf{Z}^{(\bullet)(m)}) = \widehat{\mathbf{\Psi}}_{(m)}^{(s)}(\left[\mathbf{Z}^{(1)(m)}, \ldots, \mathbf{Z}^{(s)(m)}, \ldots, \mathbf{Z}^{(S)(m)}\right]_{N\times M}) \tag{3.12}$$

where $\widehat{\mathbf{\Psi}}_{(m)}^{(s)}(\mathbf{z}^{(m)})$ denotes the estimate of $\mathbf{\Psi}_{(m)}(\mathbf{z}^{(m)})$ based on the eigenfunctions estimated from kernel PCA on the weighted gram matrix with the mth gene-set of study s:

$$\widetilde{\mathbf{K}}^{(s)} = n_s^{-1}\{w_j^{(s)} k(\mathbf{Z}_i^{(s)(m)}, \mathbf{Z}_j^{(s)(m)})\}_{1\leq i,j\leq n_s}. \tag{3.13}$$

The weight $w_j^{(s)}$ is estimated based on the sampling scheme of study s. We then estimate $\boldsymbol{\beta}^{(m)}$ with $\widehat{\boldsymbol{\beta}}^{(s)(m)}$, by constructing the pseudo-data $\{\mathbf{Y}, \widehat{\mathbf{\Psi}}_{(m)}^{(s)}\}$ where $\widehat{\mathbf{\Psi}}_{(m)}^{(s)}$ is the estimated eigenfunctions on the training data in $\mathcal{S}_{t1}$, so that $\widehat{\mathbf{\Psi}}_{(m)}^{(s)} = \{\widehat{\mathbf{\Psi}}_{(m)}^{(s)}(\mathbf{Z}^{(d)(m)})\}_{\mathcal{D}^{(d)}\in\mathcal{S}_{t1}}$ and maximizing the penalized logistic regression likelihood so that

$$\{\widehat{\boldsymbol{a}}^{(m)}, \widehat{\boldsymbol{\beta}}^{(s)(m)}\} = \underset{a,\boldsymbol{\beta}}{\operatorname{argmax}}\{\mathcal{L}(a, \boldsymbol{\beta}; \widehat{\mathbf{\Psi}}_{(m)}^{(s)})\}, \tag{3.14}$$

where

$$\mathcal{L}(a, \boldsymbol{\beta}; \widehat{\boldsymbol{\Psi}}^{(s)}_{(m)}\} = \mathbf{Y}^\intercal \log g(a + \widehat{\boldsymbol{\Psi}}^{(s)}_{(m)}\boldsymbol{\beta}) + (1-\mathbf{Y})^\intercal \log\{1 - g(a + \widehat{\boldsymbol{\Psi}}^{(s)}_{(m)}\boldsymbol{\beta}\} - \tau\|\boldsymbol{\beta}\|_2^2, \quad (3.15)$$

where $\tau \geq 0$ is a tuning parameter that can be selected via criteria such as the AIC or cross-validation, such that $n^{-\frac{1}{2}}\tau \to 0$.

We then may combine the estimates generated from using each of the training data sets in $\mathcal{S}_{t1}$ to estimate $h^{(m)}(\mathbf{z})$ by weighting the study-specific estimates with the respective sample sizes:

$$\widehat{h}^{(m)}(\mathbf{z}^{(m)}) = \frac{1}{|\mathcal{S}_{t1}|} \sum_{s \in \mathcal{S}_{t1}} \frac{n_s}{N} \widehat{\boldsymbol{\Psi}}^{(s)}_{(m)}(\mathbf{z}^{(m)}) \widehat{\boldsymbol{\beta}}^{(s)(m)} \quad (3.16)$$

with $\widehat{\boldsymbol{\Psi}}^{(s)}_{(m)}(\mathbf{z}^{(m)})$ via (3.12) and $\widehat{\boldsymbol{\beta}}^{(s)(m)}$ via (3.15). We therefore estimate $h^{(m)}(\mathbf{Z}^{(s)(m)})$ for a subject in study s in $\mathcal{S}_{t2}$ and $\mathcal{S}_v$ with genetic data $\mathbf{Z}^{(s)(m)}$ as $\widehat{h}^{(m)}(\mathbf{Z}^{(s)(m)})$. To estimate gene-set weights $\boldsymbol{\gamma}$, we construct synthetic data $\{\mathbf{Y}, \widehat{\mathbb{H}}\}$ and fit a logistic regression model as in Chapter 2:

$$\text{logitpr}(Y = 1 \mid \mathbf{Z}^{(\bullet)}) = b_0 + \boldsymbol{\gamma}^\intercal \widehat{\mathbf{H}}(\mathbf{Z}^{(\bullet)}) \quad (3.17)$$

where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_M)^\intercal$ is the vector of gene-set weights, $\widehat{\mathbf{H}}(\mathbf{Z}^{(\bullet)}) = [\widehat{h}^{(1)}(\mathbf{Z}^{(1)}), ..., \widehat{h}^{(M)}(\mathbf{Z}^{(M)})]^\intercal$ denotes the vector $\widehat{h}^{(m)}$ for all M gene-sets of an arbitrary $\mathbf{Z}^{(\bullet)}$, and $\widehat{\mathbb{H}} = \widehat{\mathbf{H}}(\mathbf{Z}^{(\mathcal{S}_{t2})(\bullet)}) = [\widehat{h}^{(1)}(\mathbf{Z}^{(\mathcal{S}_{t2})(1)}), \ldots, \widehat{h}^{M}(\mathbf{Z}^{(\mathcal{S}_{t2})(M)})]^\intercal = [\widehat{h}^{(m)}(\mathbf{Z}^{(\mathcal{S}_{t2})(m)})]_{|\mathcal{S}_{t2}| \times M}$ is the synthetic data matrix of $\widehat{h}^{(m)}$ for each subject i in a study s in the second training set $\mathcal{S}_{t2}$. We obtain a LASSO regularized estimate of $\{b_0, \boldsymbol{\gamma}\}$, as

$$\{\hat{b}, \widehat{\boldsymbol{\gamma}}\} = \underset{b, \boldsymbol{\gamma}}{\text{argmax}} \left\{\mathcal{L}_{\widehat{\mathbb{H}}}(b, \boldsymbol{\gamma}) - \tau_2\|\boldsymbol{\gamma}\|_1\right\}, \quad (3.18)$$

where $\tau_2 \geq 0$ is a tuning parameter such that $n^{-\frac{1}{2}}\tau_2 \to 0$ and $\tau_2 \to \infty$, and

$$\mathcal{L}_{\widehat{\mathbb{H}}}(b, \boldsymbol{\gamma}) = \mathbf{Y}^\intercal \log g(b + \widehat{\mathbb{H}}\boldsymbol{\gamma}) + (1 - \mathbf{Y})^\intercal \log\{1 - g(b + \widehat{\mathbb{H}}\boldsymbol{\gamma})\} \quad (3.19)$$

Note that our estimator $\widehat{\boldsymbol{\gamma}}$ is essentially an adaptive LASSO (Zou, 2006) type estimator since these weights are multiplied with $\widehat{h}^{(m)}(\mathbf{z})$ which are consistent for $h^{(m)}$; see Chapter 2 for the proof of the oracle property of $\widehat{\boldsymbol{\gamma}}$.

When analyzing several large datasets we are able to set aside a separate set of data $\mathcal{S}_{t2}$ to estimate $\boldsymbol{\gamma}$. Estimating the gene-set weights on a separate set of data $\mathcal{S}_{t2}$ than the set used to estimate $\boldsymbol{\beta}^{(m)}$ prevents overfitting of the $\widehat{\gamma}$ estimates to the nuances of the training set. If we can not afford to set aside data for this second stage of estimation, we may employ the cross-validation techniques presented in Chapter 2 to obtain the estimates for $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\gamma}$ on the same training data. It is likely that a meta-analysis would not have such sample size restrictions, however, and so we we recommend a separate $\mathcal{S}_{t2}$.

3.5 Final Prediction Model

3.5.1 Model and Algorithm

Based the estimation procedures described in previous sections, the final algorithm is as follows:

(i) Divide subjects from multiple studies into the training sets $\mathcal{S}_{t1}, \mathcal{S}_{t2}$ and the validation set $\mathcal{S}_v$.

(ii) Divide the genetic data for each subject, $\mathbf{Z}_i^{(\bullet)}$, into M gene-sets so that $\mathbf{Z}_i^{(\bullet)} = \{\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}\}$.

(iii) Estimate $\widehat{\boldsymbol{\Psi}}_{(m)}^{(s)}(\mathbf{z}^{(m)})$ via (3.12) and $\widehat{\boldsymbol{\beta}}^{(s)(m)}$ via (3.15) with each study s from training set $\mathcal{S}_{t1}$, so that $\widehat{h}^{(m)}(\mathbf{Z}^{(m)})$ is aggregated over studies in $\mathcal{S}_{t1}$ as in (3.16)

(iv) Estimate $\widehat{\boldsymbol{\gamma}}$ via (3.18) with training set $\mathcal{S}_{t2}$

(v) Use Nyström projection methods to estimate the probability of disease for a future subject with $\mathbf{z}^{(\bullet)}$ as

$$\widetilde{P}(\mathbf{z}^{(\bullet)}) = g\left[\hat{b} + \widehat{\boldsymbol{\gamma}}^{\mathsf{T}}\widehat{\mathbf{H}}(\mathbf{z}^{(\bullet)})\right] = g\left[\hat{b} + \sum_{m=1}^{M} \widehat{\gamma}_m \widehat{h}^{(m)}(\mathbf{z}^{(m)})\right]. \tag{3.20}$$

(vi) Validate the model in set $\mathcal{S}_v$ via a relevant measure of prediction accuracy, such as those described in the following section.

3.5.2 Model Validation

After such a risk prediction model is formed, it is crucial to assess its ability in discriminating subjects with or without disease. For a given risk score $\mathcal{P}$, the discrimination accuracy can be summarized based on the receiver operating characteristic (ROC) curve (Pepe, 2003). A popular measure for summarizing the overall accuracy of $\mathcal{P}$ in predicting Y is the area under the ROC curve (AUC), $\text{AUC} = \int_0^1 \text{ROC}(u)du$. We refer to Pepe (2003) for discussion about the attributes of ROC curves for evaluating diagnostic markers of disease. To obtain an accurate estimate of the AUC in a new population, we suggest withholding at least one pure validation study in $\mathcal{S}_v$ to calculate the ROC curve and AUC statistic. In this way, we do not introduce overfitting bias into our prediction accuracy estimates.

A related measure for summarizing accuracy in the presence of covariates is matched AUC (Pepe, 2000; Cai and Pepe, 2002; Dodd and Pepe, 2003). This semiparametric representation of AUC models the probability of correctly ranking the risk of individuals within a group of subjects with the same covariate value. This

is a useful method when considering genetic risk prediction adjusting for population stratification covariates. In this way, we are quantifying the ability of our prediction model to assess risk that does not depend on population stratification.

3.6 Case Study: Rheumatoid Arthritis Risk Classification with Multiple Case-Control GWAS

3.6.1 Data

Rheumatoid arthritis (RA) is the most common autoimmune inflammatory joint disease that affects 1% of the adult population worldwide (Gabriel et al., 1999). Over the past 15 years, the understanding of RA pathogenesis has advanced with the identification of environmental and genetic risk factors for the disease (Liao et al., 2009; Karlson et al., 2010). Several clinical risk factors have been implicated, such as age, gender, and smoking habits, but much of the disease liability is known to be based on genetic profiles; the narrow-sense heritability of RA has been estimated to be in the range of 0.55-0.68 (MacGregor et al., 2000; van der Woude et al., 2009).

Compiling information from a number of large scale genetic studies conducted and published in recent years, the National Human Genome Research Institute (NHGRI) provides an online catalog which lists 94 single nucleotide polymorphisms (SNPs) that have been identified as RA risk alleles (Hindorff et al., 2009, http://www.genome.gov/gwastudies/ Accessed December 10, 2011) and 90 genes that either contain these SNPs or flank the SNP on either side on the chromosome. Expanding the search to other documented autoimmune diseases (type I diabetes, Celiac disease, Crohn's disease, Lupus, Inflammatory bowel disease), the NHGRI lists

375 genes containing or flanking 365 SNPs that have been found to be associated with this class of diseases.

The SNPs that have been identified as RA risk alleles explain only a fraction of the heritability (Raychaudhuri, 2010; Stahl et al., 2010). The majority of these SNPs are located near genes of known immune functions, often in the the major histocompatibility complex (MHC) region, a large region on chromosome 6 that is known to contain a large number of genes related to immune system functions. Multiple large GWAS studies have been conducted to study RA, and several recent meta-analyses have been performed to combine many of these studies to increase power in discovering risk loci in various populations (Stahl et al., 2010; Zhernakova et al., 2011; Okada et al., 2012). With univariate testing methods, these meta-analyses have found several associated SNPs that were not previously known to be associated with RA. However, based on polygenic analysis methods, Stahl et al. (2012) estimates that at least 20% of the genetic variance of this disease is left to be found in the genome. This suggests that our blockwise methods that aim to build a prediction model with larger portions of the genome will most likely have increased accuracy over a simple additive model that merely incorporates known SNPs and genes.

The meta-data that we analyzed are drawn from 6 studies from the North America and Europe: the Brigham Rheumatoid Arthritis Sequential Study (BRASS) from the Boston area in the US, with 483 cases and 1449 controls; the CANADA study from Canada with 589 cases and 1472 controls; the Epidemiological Investigation of Rheumatoid Arthritis (EIRA) from Sweden with 1173 cases and 1089 controls; the North American Rheumatoid Arthritis Consortium (NARAC) I from North America

with 867 cases and 1041 controls; NARAC II with 462 cases and 693 controls; the Wellcome Trust Case Control Consortium from the United Kingdom with 1525 cases and 3018 controls drawn from the 1958 British Birth Cohort and the UK Blood Services. See Stahl et al. (2010) for more details about each study. All genetic data was imputed to include SNPs from the HapMap 2 release.

3.6.2 Approaches and Results

We divided the study data as follows: We chose four studies for the initial training stage $\mathcal{S}_{t1} = \{BRASS, CANADA, EIRA, NARACI\}$, one study for the second training stage $\mathcal{S}_{t2} = \{WTCCC\}$, and the last study for the final validation stage $\mathcal{S}_v = \{NARACII\}$. We also compared these results to those with $\mathcal{S}_{t2} = \{NARACII\}$ and $\mathcal{S}_v = \{WTCCC\}$. For simplicity of illustration we chose to estimate γ with a separate study in $\mathcal{S}_{t2}$, though a possible alternative approach is to utilize the cross-validation methods as presented in Chapter 2. We chose to segment the genome on the 22 autosomal chromosomes into gene-sets that include a gene and a flanking region of 20KB on either side of the gene. The data we used for analysis includes 367 gene-sets that either contain or lie up- or down-stream of the 365 SNPs that were previously found to be associated with autoimmune diseases. These 367 gene-sets cover a total of 43,345 SNPs in our dataset.

To account for any population stratification differences between populations, we include the top 5 eigenvectors from population stratification analysis in the model (Price et al., 2006), by imposing a conditional ANBKM model as follows:

$$\text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(\bullet)}, \mathbf{X}_i) = a_0 + \mathbf{X}_i^{\intercal}\mathbf{b}_0 + \sum_{m=1}^{M} \text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(m)}, \mathbf{X}_i) \qquad (3.21)$$

and

$$\text{logitpr}(Y_i = 1 \mid \mathbf{Z}_i^{(m)}, \mathbf{X}_i) = a_0^{(m)} + \mathbf{X}_i^{\intercal}\mathbf{b}_0^{(m)} + h^{(m)}(\mathbf{Z}_i^{(m)}), \tag{3.22}$$

where $\mathbf{X}_i$ represents the population eigenvectors.

Table 3.1: AUC $\times 100$ for RA risk prediction model, matched on the top five population stratification eigenvectors. The numbers in parentheses are the number of $\widehat{\gamma}$ estimates that are nonzero, or, the number of gene-sets included in the final classification rule. The columns denote the study used as a validation set.

$\mathcal{K}$	$\wp$	Block Weighting	NARAC II	WTCCC
IBS	.999	ANB	74.9 (11)	73.3 (19)
LIN	.999	ANB	75.1 (8)	73.5 (21)
IBS	.999	NB	72.9 (367)	71.1 (367)
LIN	.999	NB	69.9 (367)	69.3 (367)

Table 3.2: AUC $\times 100$ for RA risk prediction model, adjusting for G, the weighted log odds ratio risk score with known SNPs. The columns denote the study used as a validation set.

$\mathcal{K}$	$\wp$	Risk Score	NARAC II	WTCCC
IBS	.999	$h(Z) + G$	77.1	70.7
LIN	.999	$h(Z) + G$	77.4	71.4
		G	76.8	69.5

As seen in Table 3.1, the prediction accuracy in the validation set, NARAC II, is estimated with an AUC of 0.751 when using the linear kernel with 99.9% PCA, and WTCCC has an AUC of 73.5. The IBS kernel performs similarly in this data set. These AUC estimates are matched using population stratification eigenvalues from principal component analysis as covariates (Pepe, 2000; Cai and Pepe, 2002; Dodd and Pepe, 2003). This estimate allows direct comparison of risk within groups of subjects who have similar ethnic diversity. We found that matched and non-matched AUC results were similar and feel confident that population stratification has been properly accounted for in our methods.

Table 3.3: Genes (listed by their entrez ID) with non-zero estimates of $\widehat{\gamma}$ using various methods. The disease(s) that each gene has been associated with are listed in the third column. RA+ denotes that the gene has been found to be associated with RA and other autoimmune diseases.

	$\mathcal{S}_{t2}$:		WTCCC		NARAC2	
Chr	Gene (entrez)	Disease(s)	$\mathcal{K}_{\text{IBS}}$	$\mathcal{K}_{\text{LIN}}$	$\mathcal{K}_{\text{IBS}}$	$\mathcal{K}_{\text{LIN}}$
1	4774	Celiac	0.00	0.00	0.00	-1.04
1	54665	RA+	2.97	0.00	3.92	0.00
1	26191	RA+	1.45	3.94	0.00	3.57
2	5966	RA	0.00	0.00	0.00	1.58
2	150962	Cel, Chr	0.00	0.00	2.40	0.00
5	79722	RA	0.00	0.00	1.86	2.77
5	727984	RA+	0.00	0.00	1.15	0.00
6	3135	RA	0.00	0.00	0.82	1.00
6	352961	Lup	0.41	0.00	1.78	0.89
6	4277	Lup	0.00	0.00	0.00	0.06
6	55937	RA	0.00	0.00	-0.74	-0.78
6	7148	Lup	0.42	0.13	0.00	0.00
6	4855	RA+	0.39	0.17	0.72	0.15
6	10665	RA+	0.11	0.42	-0.54	-0.46
6	3122	T1D	0.50	0.81	1.93	3.16
6	3132	RA	0.56	0.00	0.00	-0.23
6	3127	RA	0.04	0.16	0.15	0.00
6	3123	RA+	0.00	0.00	0.20	0.00
6	3117	RA+	1.96	1.37	1.64	0.32
6	3119	RA+	0.45	0.82	0.69	1.26
6	3118	RA+	0.00	0.00	0.25	0.00
6	3120	Lup	0.00	0.00	0.24	0.29
7	9844	RA+	0.00	0.00	0.00	0.12
8	640	RA+	0.00	0.00	0.38	0.90
9	6366	RA	0.00	0.00	0.00	0.91
9	26147	RA+	0.00	0.00	0.00	0.27
9	7185	RA	0.00	0.00	4.02	3.04
21	1826	RA	0.00	0.00	0.66	0.95

In Table 3.3, we report the genes that have weights estimated as non-zero. Most of the genes detected have previously been found to be associated with RA, though five

previously unknown genes do appear to improve the prediction. Four of these genes lie in the MHC region on chromosome 6, while one of the genes is on chromosome 1 and has previously been discovered in studies of celiac disease. We note that prediction accuracy decreases when all gene-sets are included in the final model via the purely naive bayes method. This suggests that including noninformative blocks introduces noise into the model that decreases the prediction accuracy. Results from the LASSO in the second stage of estimation for $\widehat{\gamma}$ suggest that approximately 20 gene-sets are informative in the prediction model, while the other 340 gene-sets do not contribute to the accuracy of classification. In Table 3.2 we show the AUC values adjusted for the known genetic risk score (G) calculated from the meta-analysis log odds ratio estimates in Stahl et al. (2010). We estimated a genetic risk score based on this G alone as well as G with our estimate of risk from the NBKM meta-analysis procedures. Most of the predictive accuracy is explained by the risk score. However, including additional gene-sets in our initial model that are not known to be associated with RA may improve our ability to find signal outside of these known SNPs.

In our results we see that IBS and linear kernels perform similarly. This may be because many of the effects in these gene-sets are appropriately modeled as linear, or alternatively that an unexplored kernel may capture the effects more accurately. Further investigation is warranted and multiple kernel learning procedures may be useful to determine the best kernel to model the data.

3.7 Simulation Studies

We conducted simulation studies to determine the effects of our weighting methods on AUC in a meta-analysis setting. We simulated various case-control data sets from

a normal model and implemented our methods with an additive Gaussian kernel with varying widths. In this way, we could explicitly calculate the eigenfunctions of the kernel that we aim to estimate (Williams and Seeger, 2000). We found that the prediction accuracy gained from estimating weighted eigenfunctions is most apparent when the eigenvalues of the kernel function decays at a fairly slow rate, and the studies have fairly small sample sizes. In this case, the weighted eigenfunctions that are estimated with the small sample size span a space closer to that spanned by the true eigenfunctions than the unweighted versions. The small number unweighted eigenfunctions from the sample do not span the true eigenspace. However, as sample sizes increase and the eigenvalues decay more rapidly, the gain in prediction accuracy is less apparent because the number of weighted eigenfunctions span a similar space as the increased number unweighted eigenfunctions.

3.8 Discussion

Accurate prediction of disease outcomes with genetic markers need not be restricted by heritability that is missed by small genetic studies looking for additive effects of a small number of markers. Our gene-set meta-analysis for classification has shown that genetic susceptibility to complex diseases can be efficiently modeled via a kernel machine framework that allows for nonlinear effects of markers gene-sets, and that data from various studies with heterogenous sampling schemes can be efficiently combined to improve predictive power. Through the ANBKM model, we can achieve a balance between capturing complex effects and efficient estimation of model parameters. Incorporating the block structure of the gene-sets improve prediction accuracy and computational efficiency over global methods. By adaptively weighting the gene-

sets in the final stage of estimation, we introduce parsimony in our model that can reduce overfitting and therefore improve prediction precision for future data sets. Furthermore, we allow for the flexibility of including numerous studies with differing clinical characteristics, ancestry, and case-control sampling probabilities by controlling for any available clinical or population stratification variables, and by adapting the ANBKM method to incorporate sampling weights.

Note that when the genetic predictors are partitioned into gene-sets based on prior knowledge, we may not wish to include all gene-sets in the initial model, especially when we are partitioning over an entire genome as in this case. The original data set may include thousands of non-informative gene-sets that could unnecessary noise in our model, thereby reducing prediction performance. Therefore, we suggest implementing an initial screening procedure using a liberal threshold that removes highly non-informative gene-sets prior to building the ANBKM model. One possibility is to utilize the logistic KM score test proposed by Liu et al. (2008) that tests the null hypothesis $H_0 : h^{(m)}(\mathbf{z}) = 0$ against a general alternative, where $h^{(m)}(\mathbf{z}) \in \mathcal{H}_k$ models the effect of the gene set $\mathbf{z}$ on the outcome Y with $\mathcal{H}_k$ specified with a kernel function k, as in (3.1). This test has high power to detect a gene-set effect in high-dimensional problems because it combines signal across markers within a region and accounts for their correlation with a low degree of freedom test. It has been extended and adapted for genomic data, including genome wide sequence data, by Kwee et al. (2008) and Wu et al. (2010, 2011). Note that such a liberal screening is subsequently refined through the adaptive weighting of gene-set effects with LASSO that further reduces the number of gene-sets in the final prediction model.

Our results from the meta-analysis of RA disease risk reveal that slight predictive

power can be gained by including a number of unknown regions in the classification model. We believe that including even more unknown gene-sets would further improve accuracy. Additionally, with even more studies included in the meta-analysis, we believe that our measures of prediction accuracy will increase as we may then improve estimation of the basis functions of $\mathcal{H}_k$. The addition of clinical variables will also help account for differences in study populations and therefore allow for more accurate estimation of the contribution of genetic effects in our model. One may also consider forming sets of genetic markers based on genes or genetic pathways, and including regions in the classification model based on known pathway functions. In this way, we may uncover genes or pathways that were previously not known to be associated with the disease, but that may significantly contribute to prediction ability.

References

Bengio, Y., Delalleau, O., Le Roux, N., Paiement, J., Vincent, P., and Ouimet, M. (2004). Learning Eigenfunctions Links Spectral Embedding and Kernel PCA. *Neural Computation* **16,** 2197–2219.

Borchers, A., Uibo, R., and Gershwin, M. (2010). The geoepidemiology of type 1 diabetes. *Autoimmunity Reviews* **9,** A355–A365.

Braun, M. (2005). *Spectral Properties of the Kernel Matrix and their Application to Kernel Methods in Machine Learning.* PhD thesis, University of Bonn.

Burton, P., Clayton, D., Cardon, L., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D., McCarthy, M., Ouwehand, W., Samani, N., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678.

Cai, T., Huang, J., and Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65,** 394–404.

Cai, T. and Pepe, M. (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* **97,** 1099–1107.

Cassidy, A., Myles, J., van Tongeren, M., Page, R., Liloglou, T., Duffy, S., and Field, J. (2008). The LLP risk model: an individual risk prediction model for lung cancer. *British Journal of Cancer* **98,** 270.

Chatterjee, A. and Lahiri, S. (2010). Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society* **(accepted),**.

Chatterjee, N. and Carroll, R. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92,** 399.

Chen, J., Pee, D., Ayyagari, R., Graubard, B., Schairer, C., Byrne, C., Benichou, J., and Gail, M. (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *JNCI Journal of the National Cancer Institute* **98,** 1215.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press.

D'Agostino, R., Wolf, P., Belanger, A., and Kannel, W. (1994). Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. *Stroke* **25,** 40.

Dent, R., Trudeau, M., Pritchard, K., Hanna, W., Kahn, H., Sawka, C., Lickley, L., Rawlinson, E., Sun, P., and Narod, S. (2007). Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical Cancer Research* **13,** 4429.

Dodd, L. and Pepe, M. (2003). Semiparametric regression for the area under the

receiver operating characteristic curve. *Journal of the American Statistical Association* **98,** 409–417.

Dosaka-Akita, H., Hommura, F., Mishina, T., Ogura, S., Shimizu, M., Katoh, H., and Kawakami, Y. (2001). A risk-stratification model of non-small cell lung cancers using cyclin E, Ki-67, and ras p21: different roles of G1 cyclins in cell proliferation and prognosis. *Cancer Research* **61,** 2500.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32,** 407–451.

Eichler, E., Flint, J., Gibson, G., Kong, A., Leal, S., Moore, J., and Nadeau, J. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11,** 446–450.

Evans, D., Visscher, P., and Wray, N. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human molecular genetics* **18,** 3525–3531.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30,** 74–99.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99,** 710–723.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70,** 849–911.

Fan, J. and Peng, H. (2004). On nonconcave penalized likelihood with diverging number of parameters. *The Annals of Statistics* **32,** 928–961.

Freedman, A., Slattery, M., Ballard-Barbash, R., Willis, G., Cann, B., Pee, D., Gail, M., and Pfeiffer, R. (2009). Colorectal cancer risk prediction tool for white men and women without known susceptibility. *Journal of Clinical Oncology* **27,** 686.

Gabriel, S., Crowson, C., and O'Fallon, M. (1999). The epidemiology of rheumatoid arthritis in Rochester, Minnesota, 1955-1985. *Arthritis & Rheumatism* **42,** 415–420.

Gail, M. (2008). Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *Journal of the National Cancer Institute* **100,** 1037.

Gail, M., Brinton, L., Byar, D., Corle, D., Green, S., Schairer, C., and Mulvihill, J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81,** 1879.

Gail, M. and Costantino, J. (2001). Validating and improving models for projecting the absolute risk of breast cancer. *Journal of the National Cancer Institute* **93,** 334.

Geyer, C. (1994). On the asymptotics of constrained M-estimation. *The Annals of Statistics* **22,** 1993–2010.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13,** 135–145.

Hindorff, L., Sethupathy, P., Junkins, H., Ramos, E., Mehta, J., Collins, F., and Manolio, T. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106,** 9362.

Ioannidis, J., Ntzani, E., Trikalinos, T., Contopoulos-Ioannidis, D., et al. (2001). Replication validity of genetic association studies. *Nature genetics* **29,** 306–309.

Ioannidis, J., Trikalinos, T., Ntzani, E., and Contopoulos-Ioannidis, D. (2003). Genetic associations in large versus small studies: an empirical assessment. *The Lancet* **361,** 567–571.

Janssens, A. and van Duijn, C. (2008). Genome-based prediction of common diseases: advances and prospects. *Human molecular genetics* **17,** R166.

Jin, Z., Ying, Z., and Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88,** 381–390.

Johansen, C. and Hegele, R. (2009). Predictive genetic testing for coronary artery disease. *Critical reviews in clinical laboratory sciences* **46,** 343.

Johnson, V., Brun-Vézinet, F., Clotet, B., Conway, B., Kuritzkes, D., Pillay, D., Schapiro, J., Telenti, A., and Richman, D. (2005). Update of the drug resistance mutations in HIV-1: Fall 2005. *Top HIV Med* **13,** 125–131.

Karlson, E., Chibnik, L., Kraft, P., Cui, J., Keenan, B., Ding, B., Raychaudhuri, S., Klareskog, L., Alfredsson, L., and Plenge, R. (2010). Cumulative association

of 22 genetic variants with seropositive rheumatoid arthritis risk. *British Medical Journal* **69,** 1077.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33,** 82–95.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28,** 1356–1378.

Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6,** 113–167.

Kosorok, M. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Verlag.

Kwee, L., Liu, D., Lin, X., Ghosh, D., and Epstein, M. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics* **82,** 386–397.

Lee, S., DeCandia, T., Ripke, S., Yang, J., Sullivan, P., Goddard, M., Keller, M., Visscher, P., Wray, N., et al. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. *Nature Genetics* **44,** 247–250.

Li, H. and Luan, Y. (2003). Kernel cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing* .

Liao, K., Alfredsson, L., and Karlson, E. (2009). Environmental influences on risk for rheumatoid arthritis. *Current opinion in rheumatology* **21,** 279.

Lin, W. and Schaid, D. (2009). Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. *Genetic epidemiology* **33,** 183–197.

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics* **9,** 292–2.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics* **63,** 1079–1088.

Lohmueller, K., Pearce, C., Pike, M., Lander, E., and Hirschhorn, J. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature genetics* **33,** 177–182.

MacGregor, A., Snieder, H., Rigby, A., Koskenvuo, M., Kaprio, J., Aho, K., Silman, A., et al. (2000). Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis & Rheumatism* **43,** 30.

Machiela, M., Chen, C., Chen, C., Chanock, S., Hunter, D., and Kraft, P. (2011). Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genetic Epidemiology* **35,** 506–514.

Maher, B. (2008). The case of the missing heritability. *Nature* **456,** 18–21.

Makowsky, R., Pajewski, N., Klimentidis, Y., Vazquez, A., Duarte, C., Allison, D., and de los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. *PLoS genetics* **7,** e1002051.

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D., McCarthy, M., Ramos, E., Cardon, L., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461,** 747–753.

Marchini, J., Donnelly, P., and Cardon, L. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature genetics* **37,** 413–417.

Mardis, E. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24,** 133–141.

McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J., and Hirschhorn, J. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9,** 356–369.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: Optimality of the risk score. *Biometrics* **58,** 657–664.

McKinney, B., Reif, D., Ritchie, M., and Moore, J. (2006). Machine learning for detecting gene-gene interactions: a review. *Applied Bioinformatics* **5,** 77–88.

Meigs, J., Shrader, P., Sullivan, L., McAteer, J., Fox, C., Dupuis, J., Manning, A., Florez, J., Wilson, P., D'Agostino Sr, R., et al. (2008). Genotype score in addition to common risk factors for prediction of type 2 diabetes. *The New England Journal of Medicine* **359,** 2208.

Mukhopadhyay, I., Feingold, E., Weeks, D., and Thalamuthu, A. (2010). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic Epidemiology* **34,** 213–221.

Murcray, C., Lewinger, J., and Gauderman, W. (2009). Gene-Environment Interaction in Genome-Wide Association Studies. *American Journal of Epidemiology* **169,** 219–226.

Newey, W. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics* **4,** 2111–2245.

Okada, Y., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Kawaguchi, T., Stahl, E., Kurreeman, F., Nishida, N., et al. (2012). Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the japanese population. *Nature Genetics* .

Paynter, N., Chasman, D., Paré, G., Buring, J., Cook, N., Miletich, J., and Ridker, P. (2010). Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA: the journal of the American Medical Association* **303,** 631.

Pearson, T. and Manolio, T. (2008). How to interpret a genome-wide association study. *Journal of the American Medical Association* **299,** 1335.

Pepe, M. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association* **95,** 308–311.

Pepe, M. S. (2003). *Statistical Evaluation of Diagnostic Tests and Biomarkers.* Oxford University Press.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* , in press.

Perou, C., Sørlie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen, H., Akslen, L., et al. (2000). Molecular portraits of human breast tumours. *Nature* **406,** 747–752.

Pötscher, B. M. and Schneider, U. (2009). On the distribution of the adaptive lasso estimator. *Journal of Statistical Planning and Inference* **139,** 2775–2790.

Pötscher, B. M. and Schneider, U. (2010). Confidence sets based on penalized maximum likelihood estimators in gaussian regression. *Electronic Journal of Statistics* **4,** 334–360.

Prado, J., Wrin, T., Beauchaine, J., Ruiz, L., Petropoulos, C., Frost, S., Clotet, B., D'Aquila, R., and Martinez-Picado, J. (2002). Amprenavir-resistant HIV-1 exhibits lopinavir cross-resistance and reduced replication capacity. *Aids* **16,** 1009.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38,** 904–909.

Purcell, S., Wray, N., Stone, J., Visscher, P., O'Donovan, M., Sullivan, P., Sklar, P., Ruderfer, D., McQuillin, A., Morris, D., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752.

Rasmussen, C. (2004). Gaussian processes in machine learning. *Advanced Lectures on Machine Learning* pages 63–71.

Raychaudhuri, S. (2010). Recent advances in the genetics of rheumatoid arthritis. *Current opinion in rheumatology* **22,** 109.

Rhee, S., Gonzales, M., Kantor, R., Betts, B., Ravela, J., and Shafer, R. (2003). HIV reverse transcriptase and sequence database. *Nucleic Acids Res* **31,** 298–303.

Schölkopf, B. and Smola, A. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. The MIT Press.

Schumi, J. and DeGruttola, V. (2008). Resampling-based analyses of the effects of combinations of HIV genetic mutations on drug susceptibility. *Statistics in Medicine* **27,**.

Scott, D. (1992). *Multivariate density estimation: theory, practice, and visualization*. Wiley-Interscience.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall/CRC.

Spiegelman, D., Colditz, G., Hunter, D., and Hertzmark, E. (1994). Validation of the Gail et al. model for predicting individual breast cancer risk. *Journal of the National Cancer Institute* **86,** 600.

Stahl, E., Raychaudhuri, S., Remmers, E., Xie, G., Eyre, S., Thomson, B., Li, Y., Kurreeman, F., Zhernakova, A., Hinks, A., et al. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics* **42,** 508–514.

Stahl, E., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B., Kraft, P., Chen, R., Kallberg, H., Kurreeman, F., et al. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genetics* .

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *J. Am. Statist. Assoc.* **88,** 1350–1355.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science* **240,** 1285–1293.

Thompson, I., Ankerst, D., Chi, C., Goodman, P., Tangen, C., Lucia, M., Feng, Z., Parnes, H., and Coltman Jr, C. (2006). Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute* **98,** 529.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58,** 267–288.

Umbach, D. and Weinberg, C. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in medicine* **16,** 1731–1743.

Van Belle, T., Coppieters, K., and Von Herrath, M. (2011). Type 1 diabetes: etiology, immunology, and therapeutic strategies. *Physiological reviews* **91,** 79.

van der Woude, D., Houwing-Duistermaat, J., Toes, R., Huizinga, T., Thomson, W., Worthington, J., van der Helm-van Mil, A., and de Vries, R. (2009). Quantitative heritability of anti–citrullinated protein antibody–positive and anti–citrullinated protein antibody–negative rheumatoid arthritis. *Arthritis & Rheumatism* **60,** 916–923.

Van Marck, H., Dierynck, I., Kraus, G., Hallenberger, S., Pattery, T., Muyldermans, G., Geeraert, L., Borozdina, L., Bonesteel, R., Aston, C., et al. (2009). The Impact

of Individual Human Immunodeficiency Virus Type 1 Protease Mutations on Drug Susceptibility Is Highly Influenced by Complex Interactions with the Background Protease Sequence. *Journal of Virology* **83,** 9512.

Vasan, R. (2006). Biomarkers of cardiovascular disease: molecular basis and practical considerations. *Circulation* **113,** 2335.

Visscher, P., Hill, W., and Wray, N. (2008). Heritability in the genomics eraconcepts and misconceptions. *Nature Reviews Genetics* **9,** 255–266.

Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H., Diver, W., Thun, M., Cox, D., Hankinson, S., Kraft, P., et al. (2010). Performance of common genetic variants in breast-cancer risk models. *New England Journal of Medicine* **362,** 986–993.

Wang, H. and Leng, C. (2007). Unified LASSO estimation via least squares approximation. *Journal of the American Statistical Association* **102,** 1039–1048.

Wei, Z., Wang, K., Qu, H., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J., Chiavacci, R., et al. (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS genetics* **5,** e1000678.

Wessel, J. and Schork, N. (2006). Generalized genomic distance–based regression methodology for multilocus association analysis. *The American Journal of Human Genetics* **79,** 792–806.

Williams, C. and Seeger, M. (2000). The effect of the input density distribution on kernel-based classifiers. In *Proceedings of the 17th International Conference on Machine Learning*. Citeseer.

Williams, C. and Seeger, M. (2001). Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*. Citeseer.

Wolf, P., D'agostino, R., Belanger, A., and Kannel, W. (1991). Probability of stroke: a risk profile from the Framingham Study. *Stroke* **22,** 312.

Wray, N., Goddard, M., and Visscher, P. (2008). Prediction of individual genetic risk of complex disease. *Current opinion in genetics & development* **18,** 257–263.

Wu, M. (2009). *A parametric permutation test for regression coefficients in LASSO regularized regression.* PhD thesis, Harvard School of Public Health, Department of Biostatistics, Boston, MA.

Wu, M., Kraft, P., Epstein, M., Taylor, D., Chanock, S., Hunter, D., and Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics* **86,** 929–942.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* .

Yang, Q. and Khoury, M. (1997). Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiologic reviews* **19,** 33–43.

Yang, Q., Khoury, M., Botto, L., Friedman, J., and Flanders, W. (2003). Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *The American Journal of Human Genetics* **72,** 636–649.

Zeggini, E., Scott, L., Saxena, R., Voight, B., Marchini, J., Hu, T., de Bakker, P., Abecasis, G., Almgren, P., Andersen, G., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics* **40,** 638–645.

Zhang, H., Ahn, J., Lin, X., and Park, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22,** 88–95.

Zhang, K. and Kwok, J. (2009). Density-weighted nyström method for computing large kernel eigensystems. *Neural computation* **21,** 121–146.

Zhernakova, A., Stahl, E., Trynka, G., Raychaudhuri, S., Festen, E., Franke, L., Westra, H., Fehrmann, R., Kurreeman, F., Thomson, B., et al. (2011). Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-hla shared loci. *PLoS genetics* **7,** e1002004.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* B **67,** 301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics* **35,** 2173–2192.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36,** 1509–1533.

Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics* **37,** 1733–1751.